VNIVERSITAT DE VALÈNCIA

Facultat de Ciències Matemàtiques

Doctorat en Estadística i Optimització

Análisis y Evaluación de Hipótesis Implícitas en la Construcción de Tablas de Mortalidad

TESIS DOCTORAL

Autor:

Josep Lledó Benito

Dirigida por:

José M. Pavía Miralles

Francisco G. Morillas Jurado

Mayo 2017

Jose M. Pavía Miralles Catedrático del área de métodos cuantitativos para la Economía y la Empresa, Departament d'Economía Aplicada, Universitat de València. Francisco G. Morillas Jurado, Contratado Doctor, Departament d'Economía Aplicada, Universitat de València.

CERTIFICAN: que la presente memoria "Análisis y Evaluación de Hipótesis Implícitas en la Construcción de Tablas de Mortalidad" ha sido realizada por Josep Lledó Benito y constituye su Tesis Doctoral para optar al Grado de Doctor en Estadística.

Y para que conste, en cumplimiento de la legislación vigente, la presenta en la Facultat d'Economía de la Universitat de València, a 2 de Mayo de 2017.

Los Directores:

Jose M. Pavía Miralles

Francisco G. Morillas Jurado

A mi abuelo,

Agradecimientos

No quisiera empezar a exponer este trabajo sin antes agradecer a todas aquellas personas que me han acompañado a lo largo de estos años en esta difícil, emocionante, dura y maravillosa aventura.

A mis dos directores de tesis por su infinita paciencia. Gracias al Dr. Francisco Morillas por su ayuda y por sus numerosos consejos en los momentos más difíciles. Gracias al Dr. José M. Pavía por creer en mí antes que yo mismo lo hiciera, por su gran disponibilidad y, por supuesto, por transmitirme el conocimiento necesario para elaborar este magnífico trabajo.

A toda mi familia y amigos que son los que más sufren la gran cantidad de horas invertidas, todas ellas fuera del horario de trabajo. A mis padres y a mi abuela, por el apoyo incondicional. A mi tía Sabrina y mis primos Guillermo y Alexandre por su amor y cariño. A mis mejores amigos Albert y Héctor, grandes generadores de momentos mágicos de evasión. A mi prima Ana, por su gran apoyo y motivación.

Mi agradecimiento para Andrea por su gran comprensión y paciencia todos los días del año. A toda su familia, por sentirla como propia.

Mención especial para ti, abuelo. Contigo empecé esta tesis, te fuiste, pero volviste para ayudarme a terminarla. De ti aprendí las claves del éxito: el esfuerzo, el sacrificio y, sobretodo, la humildad.

De vosotros y para vosotros. Gracias…

Organización de la Tesis

De acuerdo con la normativa vigente (Reglament sobre depòsit, avaluació i defensa de la tesi doctoral, aprovat per el Consell de Gobern de 28 de juny de 2016, Article 8) esta Tesis Doctoral se presenta como compendio de publicaciones. No obstante, además de incluir los tres artículos publicados en revistas indexadas en algún índice internacional como JCR (WoS) y/o SJR (Scopus) también se incluye otro trabajo en proceso de evaluación en una revista indexada en los mismos índices internacionales.

Por otro lado, se incluye una introducción, una metodología, un resumen de la temática, de los principales resultados y de las conclusiones, donde se justifica la aportación original del autor.

Los trabajos incluidos en este documento son:

Introducing Migratory Flows in Life Table Construction (A1)

Statistics and Operational Research Transactions (SORT)

Assessing implicit Hypotheses in Life Tables Construction (A2)

Scandinavian Actuarial Journal

Incorporating big microdata in life table construction: A hypothesis-free estimator (A3)

Under review in Journal of Business & Economic Statistics

Trasnformations in Weekly Birth Distribution. A Temporal Analysis 1940-2010 (A4)

Revista Española de Investigaciones Sociológicas

Índice de Figuras

Índice de Tablas

Resumen

En el campo actuarial y demográfico la incidencia de la mortalidad, reflejada en la tabla de mortalidad, es utilizada en varias disciplinas. Por ejemplo, en el sistema público de pensiones o en los seguros de vida del sector asegurador. Para poblaciones generales, históricamente estas tablas han sido construidas utilizando una serie de hipótesis implícitas relativas a los distintos eventos demográficos: defunciones, migraciones y nacimientos. Sin embargo, en los últimos años hemos asistido a una explosión de información disponible, unido a una revolución en los sistemas informáticos, que mediante el desarrollo de los estimadores adecuados permite mejorar las estimaciones que actualmente se realizan en el estudio de la mortalidad.

En la construcción de la tabla de mortalidad se utilizan distintas hipótesis de manera implícita que dependen de la familia y tipo de estimadores que se empleen, y cuya necesidad depende de los niveles de información disponibles. Cuando se trabaja con estimadores de periodo AP (*period based-estimator*), basados en m_x, algunas de las hipótesis implícitas habituales consisten en asumir: (i) distribución uniforme de fallecimientos para cada edad y año, (ii) sistema demográfico cerrado o sistema demográfico abierto con hipótesis implícita de distribución uniforme de migraciones y (iii) distribución uniforme de las fechas de cumpleaños de todos los individuos de la población que no mueren a lo largo del año. Es interesante señalar que, si el estimador que se emplea está basado en el comportamiento de la cohorte durante un periodo bianual AC (*bianual-period cohort-based*), basadas en q_x, de las tres hipótesis comentadas anteriormente, la última es innecesaria.

Teniendo presente qué información está disponible en relación a la mortalidad, así como las tres hipótesis formuladas previamente, el objetivo de esta tesis es múltiple. En primer lugar, se construyen, para cada familia de estimadores, AP o AC, nuevos estimadores, cada uno de los cuales considera ninguna, alguna o varias de las hipótesis anteriores (i), (ii) y (iii). Esto, que en sí mismo ya representa una aportación en el campo de la metodología estadística, se complementa poniendo de manifiesto cómo ciertos estimadores utilizados habitualmente, en ocasiones por inercia del pasado, pueden no ser los más adecuados cuando la información a la que se tiene acceso es más detallada y se requiere cierta precisión en los resultados. Por otro lado, para medir el impacto que los estimadores propuestos pueden tener en poblaciones reales se evalúan las hipótesis

que implícitamente se asumen mediante test estadísticos de diferente naturaleza: contrastes de hipótesis espaciales, funcionales y paramétricos.

En relación a la información utilizada, en todos los casos se emplean datos oficiales, los cuales han sido proporcionados por el Instituto Nacional de Estadística (INE), en lo que corresponde a la población española, y por el Instituto Valenciano de Estadística (IVE), al abordar un estudio más detallado sobre la población de la Comunitat Valenciana. El software utilizado ha sido R y Matlab.

Abstract

In the actuarial and demographic field, the incidence of mortality, reflected in the life table, is used in different disciplines. For instance, in the public pensions system or in life insurance in the insurance sector. Historically, for general populations these tables have been built using a series of implicit hypotheses concerning different demographic events: deaths, migrations and births. However, in recent years we have witnessed an explosion of available information that, together with a revolution in information technology systems, through the development of appropriate estimators, enables improvements in the estimations that are currently made in the study of mortality.

Different hypotheses are used implicitly in the construction of the life table, depending on the family and type of estimators used; and the need for them depends on the available levels of information. When working with period estimators AP (*period based-estimator*), based on m_x, some of the usual implicit hypotheses consist of assuming: (i) uniform distribution of deaths for each age and year; (ii) closed population system or open population system with implicit hypothesis of uniform distribution of migrations; and (iii) uniform distribution of the births of all individuals who do not die throughout the year. It is interesting to note that, if the estimator used is based on the behaviour of the cohort during a biannual period AC (*biannual-period cohort-based*), based on q_x, of the three hypotheses discussed above, the latter is unnecessary.

Bearing in mind what information is available in relation to mortality, as well as the three hypotheses previously formulated, the aim of this thesis is multiple. On the one hand, new estimators are constructed for each family of estimators, AP or AC, each of which considers none, one, or several of the previous hypotheses (i), (ii) and (iii). This, which already represents a contribution to the field of statistical methodology, is complemented by showing how certain estimators customarily used, sometimes by inertia of the past, may not be the most adequate when the accessible information is more detailed and certain precision is required in the results. On the other hand, to measure the impact that the proposed estimators can have on real populations, the implicitly assumed hypotheses are evaluated using statistical tests of different natures: contrasts of spatial, functional and parametric hypotheses.

In relation to the information used, official data is utilised in all cases, provided by the Spanish Official Statistical Agency (INE), in what corresponds to the Spanish population, and the Valencian Institute of Statistics (IVE), when approaching a more detailed study of the population of the Comunitat Valenciana. The software used were R and Matlab.

Índice General

1. Introducción

En el campo actuarial y demográfico el estudio de la mortalidad es una cuestión siempre de actualidad que despierta gran interés. La mortalidad es un fenómeno natural y social que afecta a cualquier población y puede tener gran impacto en diferentes dimensiones sociales. Este hecho motiva que la evolución de la mortalidad en el tiempo (a veces en el espacio) sea considerado como un aspecto fundamental y de gran interés en áreas como (a) la planificación o evaluación del sistema de pensiones o (b) la viabilidad de un producto asegurador relacionado con los seguros de vida, por ejemplo, en los procesos de *pricing* o *reserving*.

Para el estudio de la supervivencia, las tablas de mortalidad se han consolidado como herramientas que resumen o sintetizan la información relacionada. En este sentido, cada una de estas tablas está constituida por ciertos indicadores conocidos como 'funciones biométricas', que tienen como objetivo sintetizar la mortalidad para un colectivo que posee ciertas características homogéneas; por ejemplo: un país, una región, un mismo género, una cartera de asegurados o individuos que padecen una misma patología. También, dicha información puede construirse para un momento concreto del tiempo o para estudiar un periodo de tiempo que sea corto o más extenso.

La necesidad de disponer de indicadores relacionados con la mortalidad llevó a John Graunt en 1662 a crear la primera tabla de mortalidad *o tabla de vida,* la cual fue creada utilizando el principio de Laplace. El objetivo de Graunt fue calcular las proporciones de personas que sobrevivían a cada edad, para analizar la incidencia de la peste en la población londinense de la época (Graunt, 1662). Posteriormente, Halley publicó la tabla de mortalidad de Breslau, que rápidamente usaría el gobierno británico para la comercialización de las conocidas como *rentas vitalicias*. Años más tarde, Depardieux y Cambert estimaron tablas de mortalidad para la población francesa y alemana respectivamente. Finalmente, en el siglo XIX, Gompertz (1825) y Makeham (1860) incorporan las primeras leyes de ajuste en la mortalidad para algunas edades.

No fue hasta el año 1875 cuando Wihelm Lexis introdujo la metodología para representar gráficamente el comportamiento y evolución de la mortalidad en un espacio cartesiano bidimensional, de edad *x* y año de calendario o periodo *t,* que pronto pasó a llamarse el E*squema o Diagrama de Lexis* (Lexis, 1880). Esta representación permite

mostrar la historia personal de cada uno de los individuos que componen una población, constituida por un segmento rectilíneo formando un ángulo de 45 grados con el eje temporal. En la actualidad, el estudio de la mortalidad a través del esquema de Lexis se puede realizar de forma diferente dependiendo de la información disponible en cada momento. Así, se habla de (i) modelos de un factor: cuando se considera, por ejemplo, sólo la edad; (ii) modelos de dos factores: cuando se considera la relación entre edad-periodo (AP) o la relación entre edad-cohorte (AC); y (iii) modelos de tres factores: cuando se tiene en cuenta, por ejemplo, la edad-periodo-cohorte (APC), (Booth y Tickle, 2008; Carstensen, 2007; Tabeau 2001).

En la sociedad actual el uso de la tabla de mortalidad es muy diverso y abarca diferentes disciplinas. Por ejemplo, en previsión pública es utilizada en los sistemas públicos de seguridad social para el cálculo de las pensiones públicas. En el campo demográfico la tabla de mortalidad es necesaria para generar predicciones poblacionales. En el sector asegurador y financiero se utiliza para calcular el precio de un seguro de, por ejemplo, una renta vitalicia, así como también en los planes de pensiones privados. Finalmente, su uso es también extensible a otras disciplinas como la epidemiología (ver, por ejemplo, Ahrems y Pigeot, 2007).

La tabla de mortalidad suele construirse a partir de los supervivientes a cada edad. Una manera habitual de elaborar una tabla de mortalidad es partir de un valor inicial, l_0, (habitualmente representando por 100.000 o 1.000.000 individuos) para la variable biométrica l_x, que representa el número de individuos que alcanzan la edad exacta *x*. Asumiendo la muerte como factor único de salida, dicha variable disminuye edad a edad a causa de la incidencia de la mortalidad en el colectivo; representada por la probabilidad de que una persona de edad *x* fallezca antes de alcanzar la edad *x+1*, q_x. También es habitual utilizar la variable biométrica, m_x, definida como el cociente entre el número de personas que fallecen a la edad *x* y el número medio de personas que están expuestas al riesgo en el periodo *t*.

Las tablas de mortalidad son construidas por diferentes entidades u organismos oficiales. Así, para una población general, las tablas de mortalidad de España son construidas por el Instituto Nacional de Estadística (INE), mientras que en Reino Unido lo son por The Office for National Statistics. Por otro lado, en la población asegurada,

esta labor puede ser realizada por el regulador de cada país, en España la Dirección General de Seguros (DGS), o por las propias compañías aseguradoras.

Generalmente, las tablas de mortalidad utilizadas por los agentes económicos públicos y privados son construidas utilizando hipótesis implícitas que dependen de la información disponible y del estimador empleado. Así, cuando se trabaja con estimadores de periodo, *period based*-estimator o modelo edad-periodo (en adelante, *AP*), centrados en m_x, (ver INE 2016; ONS, 2010, 2012 y Wilmoth *et al.*, 2007), algunas de las hipótesis implícitas más habituales consisten en asumir (H1) distribución uniforme de fallecimientos para cada edad *x* y año *t*; (H2i) sistema demográfico cerrado (o al menos no consideración explícita de los flujos migratorios); (H2ii) sistema demográfico abierto aunque distribución uniforme de los flujos migratorios y (H3) distribución uniforme de los cumpleaños de los individuos de la población que no mueren a lo largo del año. Si el estimador empleado está basado en el comportamiento de la cohorte durante un periodo bianual, bianual *cohort based-estimator* o modelo edad-cohorte (en adelante, *AC*), centrados en q_x (INE 2007), de las tres hipótesis comentadas anteriormente, la última es innecesaria. Para ambas familias de estimadores se asumen otras dos hipótesis adicionales de manera implícita; (H4) que los inmigrantes adquieren el mismo riesgo que la población residente y, (H5) que aquellos individuos que emigran tienen similar riesgo de fallecer que la población que permanece.

Adicionalmente a las diferencias en las hipótesis entre los dos estimadores, estos también difieren en la variable biométrica en la que se centran para obtener el resto de componentes de la tabla de mortalidad. Así, el estimador AC, basado en la cohorte, se utiliza para obtener la probabilidad de fallecimiento, q_x, mientras que el estimador AP, basado en el periodo, se utiliza para obtener la tasa de fallecimiento, m_x. En la actualidad, no parece apreciarse un consenso sobre el estimador a utilizar, ya que ambas variables biométricas son el punto de partida para el cálculo posterior de la tabla de mortalidad. Sin embargo, las tendencias actuales (INE 2016; ONS, 2010, 2012; Arias 2015) parecen decantarse antes por emplear estimadores de la familia AP frente a estimadores de la familia AC (INE, 2007).

1.1. Motivación

Históricamente las tablas de mortalidad han sido construidas con datos agregados relativos al número de defunciones por edad y sexo, y se han omitido de manera implícita tanto los flujos migratorios como los momentos exactos de defunción y de nacimiento. Sin embargo, en los últimos años hemos asistido a dos procesos relevantes que están teniendo una enorme repercusión en el campo estadístico, especialmente en la estadística operativa y en la optimización. Por un lado, estamos asistiendo a una recopilación masiva de información o *big data,* observable también en el campo demográfico (Ruggles, 2014), y a su posterior tratamiento a través de procesos automáticos de tratamiento de datos. Por otro lado, estamos viviendo una gran revolución tecnológica (IT) que permite el tratamiento masivo de información en ordenadores personales. Estos acontecimientos han propiciado que recientemente se hayan observado algunos avances en la construcción de las tablas de mortalidad realizadas por los agentes oficiales. Por ejemplo, el INE ha introducido el momento exacto del fallecimiento en sus estimaciones oficiales de mortalidad (INE, 2009).

A pesar de estos pequeños avances en materia demográfica y estadística operativa, el resto de hipótesis (H2ii) y (H3), comentadas en la introducción, no han sido incorporadas aún en las estimaciones de probabilidades y tasas de fallecimiento, q_x y m_x, oficiales. Sería de enorme interés en el campo estadístico-actuarial disponer de estimadores que permitiesen incorporar dichas hipótesis. También, sería interesante conocer el impacto en términos económicos de su introducción en las disciplinas en las que la tabla de mortalidad es utilizada. Por este motivo, las preguntas de investigación que motivan la realización de esta tesis son: (i) ¿Qué sucede y cómo se definen estimadores de tipo AP y AC al no considerar una o varias de las hipótesis restrictivas comentadas?; (ii) ¿Es recomendable asumir estas hipótesis cuando se disponen de datos más detallados?; (iii) ¿Qué diferencias se producen y cuál es el impacto de las mismas de emplear los estimadores que se proponen frente a los estimadores denominamos *clásicos*, tanto en el sistema público de pensiones como en determinados productos aseguradores? (iv) ¿Cómo han sido tratadas estas hipótesis en otras disciplinas relacionadas y cuál es su impacto?

Al respecto de la pregunta (i) el desarrollo para eliminar cada hipótesis se realiza de manera paulatina. Así, se parte del estimador AP o AC que contenga todas las hipótesis implícitas y de manera gradual se van eliminando cada una de ellas. Este proceso nos permite poder analizar el impacto en las tablas de mortalidad al no considerar cada hipótesis en los estimadores.

La respuesta a la pregunta (ii) requiere un análisis pormenorizado de la población objetivo de estudio. Por ejemplo, considerar la hipótesis de sistema demográfico cerrado no parece adecuado en un país con un elevado flujo migratorio. Hasta el año 2008 (ver Figura 43-SA3) España se caracterizó por ser un país con un flujo migratorio positivo (el número de inmigrante superaba al número de emigrantes). A partir de dicha fecha, y como consecuencia de la situación económica, miles de jóvenes e inmigrantes previos abandonaron el país con el objetivo de buscar oportunidades profesionales en otros mercados. Sin duda, estos acontecimientos condicionan enormemente la población expuesta al riesgo susceptible de sufrir el fallecimiento y, por lo tanto, es esperable que la incorporación de los flujos migratorios (H2) afecte a las estimaciones de mortalidad. Siguiendo en la misma pregunta de investigación, podemos esperar que la hipótesis de distribución uniforme de nacimientos tampoco se cumpla. Por ejemplo, la presencia de eventos históricos como la Guerra Civil española ocasionó una concentración (escasez) de nacimientos en los meses posteriores a (durante) la contienda.

Es de esperar que si las hipótesis anteriores no se cumplen la respuesta a la pregunta (iii) sea que sí existen diferencias. Así, las probabilidades y tasas de fallecimiento obtenidas al utilizar cada uno de los estimadores pueden tener un impacto no despreciable en el sistema de pensiones y en los productos de seguros de vida. Por ejemplo, es importante notar el incremento continuado en el pago que realiza Seguridad Social en concepto de prestaciones (pensiones) a la población pasiva (pensionistas). Las modificaciones en las tasas de mortalidad o probabilidades de fallecimiento pueden ocasionar diferencias sustanciales en la partida de gasto más importante en nuestro país (38,5% del total de la partida de gastos en el año 2016). Por otro lado, según los datos económicos de la Investigación Corporativa entre Entidades Aseguradoras (ICEA), en 2015 las aseguradoras españolas tenían en el pasivo del balance contable un total de

69.959,91[1] millones de euros en concepto de reservas o provisiones técnicas de seguros de vida. De nuevo, una modificación de la tabla de mortalidad puede convertirse en un problema de gran importancia que afecte a una proporción importante de compañías aseguradoras.

[1] 51,635.48 millones de euros correspondientes al negocio individual, 14,999.75 millones de euros en rentas en fase de cobro instrumentados en compromisos por pensiones y 3,324.68 millones de euros en rentas en fase de cobro en seguros colectivos.

1.2. Objetivo

La actual estructura demográfica poblacional supone un reto para nuestros agentes económicos. Es importante destacar que las generaciones nacidas en el periodo 1965-1979, la que se conoce como *Generación X*, pronto se incorporarán al colectivo pasivo de la sociedad. Este acontecimiento supone un incremento importante en forma de prestaciones sociales (pensiones) que debe de ser tenido en cuenta para calcular las cotizaciones que se deben de realizar para garantizar la viabilidad del sistema actual de jubilación. Por otro lado, el sector asegurador comercializa productos destinados a cubrir tanto pensiones privadas (complementarias a las pensiones públicas) como a asegurar la propia vida. De esta manera queda evidenciado como, tanto en el ámbito público como en el privado, la estimación de la mortalidad futura es algo necesario y que requiere cierta precisión. Por este motivo, la utilización de ciertas hipótesis en la definición de los estimadores empleados en caracterizar la mortalidad puede propiciar desviaciones en las estimaciones realizadas y propagarse hacia las predicciones de la mortalidad futura.

En esa línea, el objetivo de la tesis doctoral ha sido doble. Por un lado, se han desarrollado y construido los estimadores de tipo AC y AP eliminando, paso a paso, cada una de las hipótesis implícitas utilizadas para la construcción de las tablas de mortalidad. Lo cual supone disponer de nuevos y valiosos estimadores en este ámbito. Por otro lado, la evaluación de las hipótesis implícitas se ha realizado haciendo uso de contrastes de hipótesis espaciales, funcionales y paramétricos con el objetivo de analizar la idoneidad de las mismas. Asimismo, el desarrollo y la evaluación de las distintas hipótesis ha permitido estudiar el impacto de las mismas en otras disciplinas específicas de las ciencias sociales.

Las investigaciones desarrolladas que han dado lugar a esta tesis doctoral y que dan respuesta a las múltiples preguntas de investigación planteadas han sido sintetizadas en cuatro artículos científicos titulados: (A1) *Introducing migratory flows in life table construction*, (A2) *Assessing implicit hypotheses in life table construction*, (A3) *Incorporating big microdata in life table construction: A hypothesis-free estimator, y* (A4) *Transformations in Weekly Birth Distribution. A Temporal Analysis 1940-2010.* Los dos primeros están centrados en el estimador de tipof AC, mientras que el tercer trabajo se

centra en el análisis del estimador de tipo AP. Por otro lado, el cuarto trabajo tiene como objetivo profundizar en el origen que tiene el rechazo generalizado de la hipótesis (H3) en el campo de las ciencias sociales. Para ello, se estudia la estacionalidad de los nacimientos y sus cambios a lo largo de las últimas décadas alertando de los posibles efectos de asumir dicha hipótesis en la ciencia actuarial y demográfica.

1.3. Datos y Software

Para el desarrollo de cada uno de los artículos ha sido necesaria la utilización de bases de datos para un conjunto de años y diferentes poblaciones. En concreto, para el trabajo (A1) se utiliza información, tanto de la población residente (el número de personas con vida por edad *x*, en el periodo 2006-2008) como de las variaciones residenciales[2] (el número de emigrantes e inmigrantes por edad x, y para cada año en el periodo 2006-2008), disponible en el sitio web del INE (http://www.ine.es. Para la realización del trabajo (A2) el requerimiento de información ha sido mayor. Además de los ficheros utilizados en el trabajo anterior, se solicitó al INE el fichero de micro-datos pertenecientes a las defunciones durante el mismo periodo analizado, 2006-2008.

El tercer trabajo (A3) es sin duda el que mayor información detallada ha requerido. Los micro-datos utilizados en el artículo pertenecen a la población de la Comunitat Valenciana, más de 20 millones de habitantes para los 4 años analizados. En concreto, se han utilizado micro-datos para cada sexo correspondientes a las fechas de defunción, a las fechas de inmigración y emigración y a las fechas de cumpleaños de todos los individuos que componen la población de los años 2010 a 2013. Los micro-datos de fallecidos (día, mes y año de defunción y nacimiento) y de flujos migratorios (día, mes y año de migración y nacimiento) han sido proporcionados por el INE. Los datos correspondientes a los padrones de población (día, mes y año de nacimiento) han sido proporcionados por el Instituto Valenciano de Estadística (IVE), correspondiéndose consistentemente con los datos oficiales del INE. De igual modo, los datos empleados para el desarrollo del artículo (A4) provienen del padrón municipal de habitantes de la Comunitat Valenciana para el año 2010, proporcionados también por el IVE.

Los datos pertenecientes a los inmigrantes utilizados en los trabajos A2 y A3 han sido tratados previamente. Fuimos advertidos por los agentes estadísticos oficiales del INE que en el fichero de variaciones residenciales cuando se desconoce la fecha de cumpleaños de un inmigrante, administrativamente se establece el 1 de enero. Para mitigar este exceso de nacimientos generado artificialmente se ha analizado en ambos trabajos la distribución del número de inmigrantes a lo largo de los días de cada año y

[2] http://www.ine.es/prodyser/micro_varires.htm

se ha comparado con el 1 de enero. El exceso se ha repartido aleatoriamente durante todos los días del año.

Los cálculos han sido realizados en su gran mayoría con el software estadístico libre R (R Core Team, 3.0.2 y 3.3.0 de 2013, 2016) y con el software matemático de propósito general Matlab. También, algunas de las imágenes se realizaron con Microsoft Office Excel[3]. Adicionalmente, una serie de librerías contenidas en el software R han sido utilizadas en el apartado de evaluación de hipótesis implícitas en los trabajos (A2) y (A3): el package *spatstat* (Baddeley and Turner 2005) y el package *GoFKernel* (Pavia, 2015).

[3] La variedad del software utilizado se debe a cuestiones de conocimiento previo y evolución científica. Inicialmente el software utilizado ha sido Matlab, y se ha finalizado esta tesis haciendo uso intensivo del software libre R. Es importante señalar que ambos programas son ampliamente utilizados en el ámbito académico, robustos y fiables en cuanto a los resultados que proporcionan. De esta manera, y dado que cada software ha sido aplicado en un artículo diferente, no pensamos que los resultados obtenidos en los trabajos realizados hubieran dado lugar a conclusiones diferentes.

2. Metodología

2.1. El Esquema de Lexis

La derivación matemática de los estimadores propuestos en la tesis parte de representar cada movimiento demográfico en el Esquema de Lexis, utilizando cada uno de los micro-datos disponibles (Figura 1). En concreto, el esquema de Lexis permite representar gráficamente, en un espacio cartesiano bidimensional de edad *x* y año de calendario o periodo *t*, distintos acontecimientos demográficos. La base de la representación es mostrar cada uno de los individuos que componen una población mediante un segmento rectilíneo formando un ángulo de 45 grados con el eje temporal. En el esquema de Lexis cada historia personal comienza con el nacimiento (líneas con inicio en la base del diagrama), o con la inmigración, denotada por 'o' (líneas e y d), y termina en el fallecimiento, denotada por 'x' (líneas a y b), o por la emigración, denotada por '□' (línea c) (Willmoth *et al*. 2007). En una sección cualquiera del esquema de Lexis las personas que continúan con vida con edad *x* al final del año *t* también son tenidas en cuenta (líneas f y g).

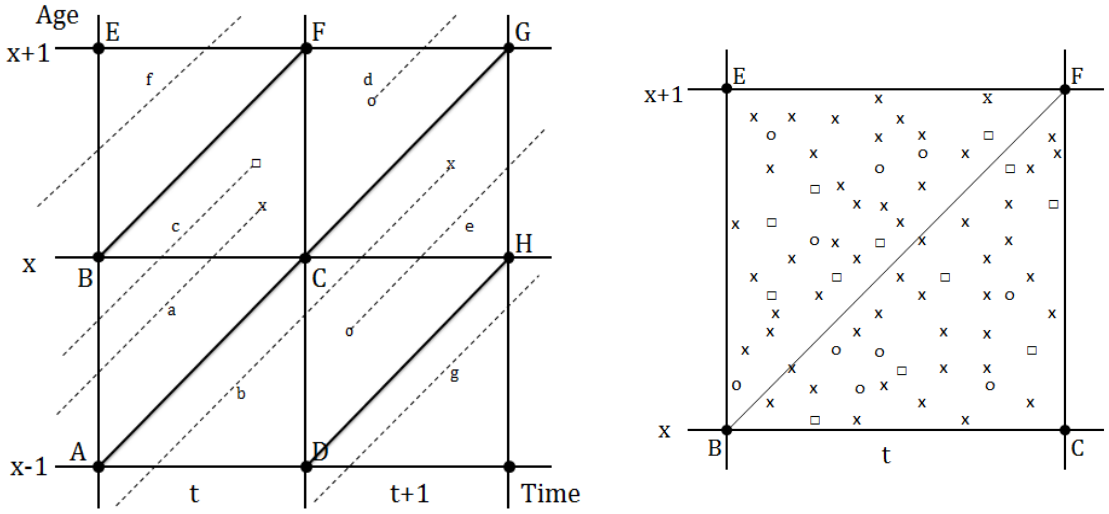

Figura 1. Pequeña sección del diagrama de Lexis con algunas líneas de vida (izquierda) y representación esquemática de los eventos de fallecimiento (×) y migración (□, ○) en una celta 1x1 (derecha).

Otra de las abstracciones del esquema de Lexis consiste en asignar valores a las superficies que lo forman. La longitud del segmento BE (CF) representa el número de personas con vida con edad *x* el 1 de enero del año *t* (*t+1*), denotado por C_x^t (C_x^{t+1}). El número de defunciones se corresponde con el total de cruces que finalizan su línea de vida en el área BCEF, denotada por D_x^t. A su vez, se denota como $D_{x:t-x}^t$ el número de personas que han fallecido con edad *x* cumplida a lo largo del año *t* (cruces que finalizan

en el triángulo inferior BCF). De igual modo, $D^{t}_{x:t-1-x}$ se corresponderá con el número de personas que han fallecido con edad *x* cumplida a lo largo de *t-1* (cruces que finalizan en el triángulo superior BEF).

Respecto a los flujos migratorios, el número de cuadrados (círculos) que finalizan (comienzan) en la superficie BCEF se corresponde con el número de emigrantes (inmigrantes) que emigran (inmigran) con edad *x* en el año *t*, denotado por E^{t}_{x} (I^{t}_{x}). A su vez, se denota como $E^{t}_{x:t-x}$ ($I^{t}_{x:t-x}$) el número de personas que han emigrado (inmigrado) con edad *x* cumplida a lo largo del año *t* (cuadrados y círculos que finalizan y empiezan en el triángulo inferior BCF). Por otro lado, $E^{t}_{x:t-1-x}$ ($I^{t}_{x:t-1-x}$) se corresponderá con el número de personas que han emigrado (inmigrado) con edad *x* cumplida a lo largo de *t-1* (cuadrados y círculos que finalizan y empiezan en el triángulo superior BEF).

Cuando se quiere medir la distancia entre dos puntos de un evento demográfico acontecido en un área del esquema de Lexis, nuevas variables deben ser introducidas (Figura 2). Así, $b^{t}_{x,j}$ y $e^{t}_{x,O}$ (segmento JK y segmento OQ), se definen como el tiempo transcurrido en años entre la fecha de defunción (emigración) y la fecha de cumpleaños de edad *x*, de los individuos que fallecen (emigran) con edad *x* y que cumplen años a lo largo de *t* (cruces o cuadrados que finalizan en el área que forma el triángulo inferior BCF). Para aquellos individuos que fallecieron (emigraron) con edad *x* en el año *t* pero que cumplieron años a lo largo de *t-1*, dicha cantidad, $b^{t}_{x,L}$ y $e^{t}_{x,X}$, es negativa y coincide con la distancia entre la fecha de defunción (emigración) y la recta de cumpleaños de edad *x+1*, (segmento LN y segmento XS). Se corresponde con las cruces o cuadrados que finalizan en el área que forma el triángulo superior, BEF.

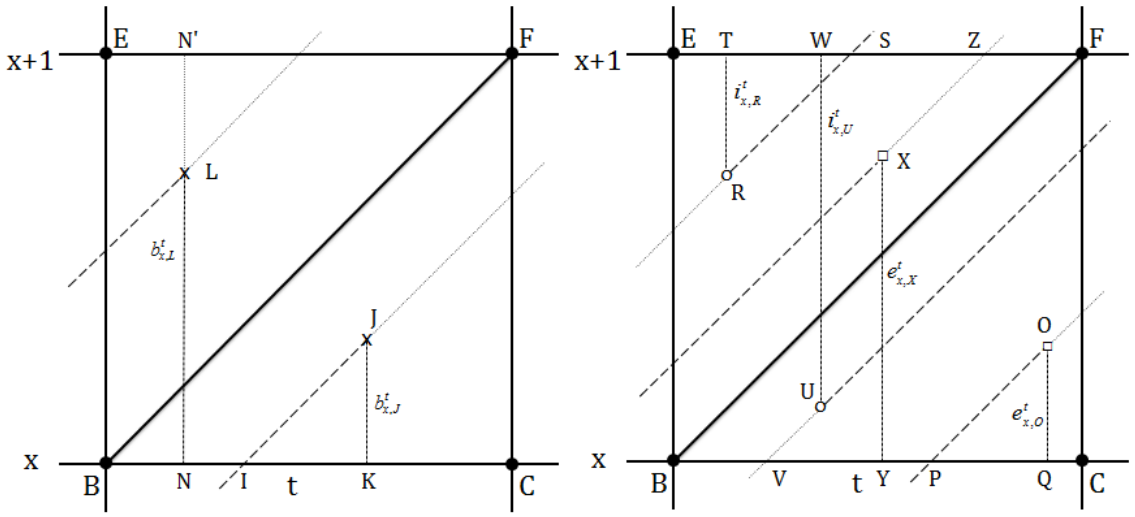

Figura 2. Detalle en el esquema de Lexis para el fallecimiento (izquierda) y eventos migratorios (derecha). $b^t_{x,L}$, $b^t_{x,J}$ $e^t_{x,X}$ y $e^t_{x,O}$ miden la distancia entre el evento de fallecimiento (L y J) o inmigración (X y O) de los individuos correspondientes y la fecha de su x cumpleaños. Excepto en el excepcional caso (como cuando una persona inmigra y emigra o emigra e inmigra con la misma edad), estas cuantías representan el tiempo exacto (en años) expuesto al riesgo de morir con la edad completada x como miembro(s) de la población objetivo para los individuos identificados por los puntos J, L, R, X, U y O.

Para el caso de los inmigrantes, $i^t_{x,U}$ se define como la diferencia en años entre la fecha de inmigración y la fecha de cumpleaños de edad *x*. Dicha cantidad es negativa (segmento UV) para la generación que cumplió *x* años a lo largo de *t* (círculos que empiezan en el área que forma el triángulo inferior BCF, punto U) y es positiva (segmento RT) para la generación que cumplió *x* años a lo largo de *t-1* (círculos que empiezan en el área que forma el triángulo superior BEF, punto R). En esa línea, se han creado otras variables necesarias para desarrollar y crear los distintos estimadores, de tipo AP y AC, presentados en la introducción del presente trabajo. Para una mayor comprensión de dichas variables se pueden consultar las tablas: Table 1 del artículo (A2) y Table 1 del artículo (A3), así como su descripción en detalle en las segundas secciones de los respectivos artículos.

Para medir la intensidad de la mortalidad en el colectivo a través de un estimador se utiliza una función biométrica apropiada a cada tipo de estimador (Edad-Periodo o Edad-Cohorte). (i) En el caso de un estimador AC se utiliza la probabilidad que una persona de edad cumplida *x* no alcance la edad *x+1*, q_x (artículos A1 y A2). Para su obtención se divide el número de fallecidos de edad cumplida *x*, nacidos en la cohorte *t-x*, entre la población existente al inicio del periodo *t* con esa edad, la población expuesta al riesgo. Para recoger el efecto cohorte se utilizan datos de dos periodos

consecutivos, *t* y *t+1*. Así, una persona de edad *x* (nacida en la cohorte *t-x*) puede fallecer con edad *x* en el año *t* o con edad *x* en el año *t+1* (antes de su próximo cumpleaños). (ii) En el caso de un estimador AP se obtiene la tasa de fallecimiento de edad *x*, m_x (artículo A3), como el cociente entre el número de fallecidos acontecidos con edad cumplida *x* y la población media expuesta al riesgo (o el "número de años" totales expuestos al riesgo por la población objeto de estudio) durante el mismo periodo *t*. En este caso el estimador se construye en relación a información de un solo año, perdiendo la cohorte relevancia en el cálculo. Esto queda claro si se piensa que el numerador solo tiene en cuenta el número de fallecidos con edad cumplida *x* en el año *t*, con independencia de a qué cohorte pertenece.

Como es habitual en el cálculo actuarial y demográfico, las probabilidades o tasas de fallecimiento brutas obtenidas pueden presentar variaciones bruscas de unas edades a otras. Al tratarse de un proceso estocástico, la evolución de las probabilidades o tasas brutas de fallecimiento no siempre se corresponden con el comportamiento esperado de la mortalidad. Con el objetivo de suavizar los movimientos bruscos presentes en dichos datos empíricos se pueden utilizar diversos procedimientos. En los artículos que componen esta tesis se ha utilizado un método de ajuste o graduación no paramétrico. En concreto, se ha utilizado el modelo gaussiano no paramétrico Kernel con parámetro ventana o *bandwidth* igual a 2 (Ayuso *et al.*, 2007).

La incorporación en el cálculo de la probabilidad y la tasa de fallecimiento de cada una de las hipótesis implícitas se realiza de manera paulatina para cada uno de los dos estimadores. Así, se parte de un estimador AP o AC que contiene todas las hipótesis implícitas en su construcción y se eliminan paso a paso las tres hipótesis comentadas para el estimador AP y las dos primeras hipótesis para el estimador AC. En concreto, el trabajo A1 tiene como objetivo eliminar la hipótesis (H2i) al introducir los movimientos migratorios agregados en el cálculo de la tabla de mortalidad en un estimador AC. Profundizando en este estimador, el trabajo A2 amplía el trabajo desarrollado por el estimador A1 e incorpora los momentos exactos de las defunciones (H1) y de las migraciones (H2ii) en las estimaciones de la probabilidad de fallecimiento junto con una evaluación exhaustiva de todas las hipótesis. Finalmente, el trabajo A3 tiene como objetivo realizar el mismo desarrollo estadístico, pero en el marco de un estimador AP.

2.2. Test de hipótesis

En este apartado se introducen los test estadísticos utilizados para evaluar las hipótesis de distribución uniforme de fallecidos (H1), la hipótesis de sistema demográfico cerrado (H2i), la hipótesis de sistema demográfico abierto, la hipótesis de distribución uniforme de migraciones (H2ii) y la hipótesis de distribución uniforme de nacimientos, utilizando tres tipos de contrastes: espaciales, funcionales y paramétricos.

En el apartado de evaluación de hipótesis implícitas miles de test estadísticos han sido realizados. En concreto, para el estimador AP se han evaluado las tres hipótesis utilizadas (H1, H2i, H2ii y H3), en ambos sexos, y en las tres figuras geométricas del esquema de Lexis (cuadrado, triángulo interior y triángulo superior) para los cuatro años comprendidos entre 2010 y 2013. Por otro lado, para el estimador AC se han analizado las dos primeras hipótesis (H1, H2i y H2ii), ambos géneros y cuatro figuras geométricas (cuadrado, triángulo interior, triángulo superior y cuadrilátero) para los tres años comprendidos entre 2006 y 2008. En total, más de 98.000 test se han programado mediante scripts ad-hoc en el software R.

Para representar esta gran cantidad de test se han creado gráficos de rejillas que sintetizan de manera original el resultado de los distintos p.values por edad, sexo y tipología de contraste. En ese sentido, a lo largo de los trabajos (A2) y (A3) se exponen varios ejemplos. Adicionalmente, se han decidido crear dos supplementary material (SA2 y SA3) con el objetivo de no sobrecargar de figuras los trabajos realizados y ofrecer al lector interesado toda la información.

Contrastes espaciales

Desde un punto de vista geométrico, el conocimiento de las fechas de cada evento demográfico permite ubicar con exactitud defunciones (cruces), inmigraciones (cuadrado) y emigraciones (círculos) en cada figura geométrica (cuadrado, triángulo inferior, triángulo superior y cuadrilátero) del diagrama de Lexis y contemplar los datos como procesos puntuales. Una vez contemplados los conjuntos de puntos como realizaciones de un patrón de puntos bivariable en un espacio Cartesiano (año, edad) es posible testar la hipótesis de Complete Spatial Randomness (CSR), que es el equivalente espacial de la hipótesis de uniformidad.

La hipótesis de CSR ha sido probada usando algunos test disponibles en el package *spatstat* (Baddeley and Turner 2005). En particular, hemos usado el CLF test (Cressie 1991; Loosmore and Ford 2006), el Maximum Absolute Deviation (MAD) test (Ripley 1977, 1981) y una versión del ya conocido Chi-squared goodness-of-fit (XS) test. Para el último test se han divido los triángulos y cuadrados en ocho y dieciséis partes respectivamente, y se midió la distancia de Chi-Cuadrado de Pearson entre el número de puntos observado y el número de puntos esperado.

Contrastes funcionales

Adicionalmente a los test espaciales, hemos utilizado también dos test funcionales para evaluar una serie de hipótesis. Por un lado, en el trabajo (A2) se ha evaluado (i) si el tiempo vivido por aquellos individuos que fallecen con edad cumplida *x* se distribuye como una variable aleatoria uniforme en el intervalo [0,1]. De igual modo, (ii) se ha analizado la distribución del tiempo de exposición de inmigrantes y emigrantes. Otro análisis realizado ha consistido (iii) en medir el tiempo de exposición al riesgo de fallecidos, inmigrantes y emigrantes en cada uno de los triángulos. Por otro lado, en el trabajo (A3) el objetivo ha consistido en contrastar la compatibilidad entre la distribución teórica de tiempos de exposición (no-exposición) que se deriva de las hipótesis de uniformidad y las distribuciones empíricas de tiempo de exposición (no-exposición) al riesgo que se obtienen en cada triángulo para cada tipo de evento demográfico.

Para estos análisis se han utilizado el test de Kolmogorov–Smirnov (KS) (ver, por ejemplo, Conover, 1971) programado en la función ks.test y el test Geometric (G) disponible en la librería *GoFKernel* (Pavía, 2015).

Contrastes paramétricos

Finalmente, se ha utilizado también una batería de test paramétricos, pues en las fórmulas de los estimadores las hipótesis implícitas se concretan en coeficientes específicos. En el trabajo (A2) examinamos si (i) el número de defunciones ocurridas en ambos triángulos son iguales por edad y sexo; (ii) el número de inmigrantes y emigrantes son iguales también en ambos triángulos; (iii) el tiempo medio expuesto al riesgo de los fallecidos localizados en el triángulo inferior (superior) es $\frac{2}{3}$ ($\frac{1}{3}$); (iv) el tiempo medio

expuesto al riesgo de los emigrantes (inmigrantes); y (v) si el tiempo medio vivido por los individuos fallecidos con edad *x* es $\frac{1}{2}$. Por otro lado, en el trabajo (A3) evaluamos si el tiempo medio de exposición (no-exposición) al riesgo de las personas que mueren o migran es $\frac{1}{3}$ en ambos triángulos. Por otra parte, estudiamos si el número de fallecidos, emigrantes e inmigrantes es igual en los dos triángulos. Los tests empleados son t-tests de medias y el tests binomiales (con $p = 0.5$).

3. Estimador de dos factores

3.1. Estimador basado en la cohorte

En el proceso de desarrollar un estimador AC es posible tener en cuenta dos hipótesis implícitas. La hipótesis de distribución uniforme, de fallecidos (H1) y migrantes (H2ii), y la hipótesis de sistema demográfico cerrado (H2i). Este estimador parte del stock de población a 1 de enero del año *x-1* de la población objetivo de estudio, $\mathrm{C}_{\mathrm{x}-1}^{\mathrm{t}}$. Para cada escenario, edad y sexo, a dicha cantidad se le añade el número de personas susceptibles de sufrir el fallecimiento, la población inicial y los emigrantes, y se detrae el número de personas que lo han (no tienen la posibilidad) sufrido (de sufrirlo), las defunciones (los inmigrantes). Las expresiones recogidas en la Tabla 1 permiten calcular de manera sencilla el estimador correspondiente para cada nivel de información.

Table 1. Resumen de los estimadores de un cohort-based estimator.

Estimador	Hipótesis		
	Sistema demográfico Cerrado (CP)	Uniforme	
		Fallecidos (D)	Migraciones (M)
$\hat{q}_x = \frac{\frac{1}{2}(D_x^t + D_x^{t+1})}{C_{x-1}^t - \frac{1}{2}D_{x-1}^t}$	Si	Si	N/A
$\tilde{q}_x = \frac{\frac{1}{2}(D_x^t + D_x^{t+1})}{C_{x-1}^t - \frac{1}{2}D_{x-1}^t + \frac{1}{2}N_{x-1}^t + \frac{1}{3}N_x^t + \frac{1}{6}N_x^{t+1}}$	No	Si	Si
$\breve{q}_x = \frac{D_{x:t-x}^t + D_{x:t-x}^{t+1}}{C_{x-1}^t - D_{x-1:t-x}^t} = \frac{D_{x:t-x}^{t,t+1}}{C_{x-1}^t - D_{x-1:t-x}^t}$	Si	No	N/A
$\ddot{q}_x = \frac{D_{x:t-x}^t + D_{x:t-x}^{t+1}}{\ell_x^t - E_{x:t-x}^{t,t+1} + \sum_{j=1}^{E_{x:t-x}^t} e_{x,j}^t + \sum_{j=1}^{E_{x:t-x}^{t+1}} e_{x,j}^{t+1} + \sum_{j=1}^{I_{x:t-x}^t} i_{x,j}^t + \sum_{j=1}^{I_{x:t-x}^{t+1}} i_{x,j}^{t+1}}$	No	No	No

$\ell_x^t = C_{x-1}^t - D_{x-1:t-x}^t + N_{x-1:t-x}^t$ and $R_x^t = \ell_x^t - E_x^{t,t+1} + \sum_{j=1}^{E_{x:t-x}^t} e_{x,j}^t + \sum_{j=1}^{E_{x:t-x}^{t+1}} e_{x,j}^{t+1} + \sum_{j=1}^{I_{x:t-x}^t} i_{x,j}^t + \sum_{j=1}^{I_{x:t-x}^{t+1}} i_{x,j}^{t+1}$. N/A: No aplica.

El estimador AC que menos información necesita es el que asume sistema demográfico cerrado y distribución uniforme de fallecidos (CDS_UD). Así, la probabilidad de que una persona de edad *x* no alcance la edad *x+1* se obtiene como sigue; $\hat{q}_x = \frac{\frac{1}{2}(D_x^t + D_x^{t+1})}{C_{x-1}^t - \frac{1}{2}D_{x-1}^t}$. El segundo estimador permite abrir el sistema demográfico, aunque considera distribución uniforme de fallecidos y migraciones (ODS_UDM). Por lo tanto, al estimador anterior se le añade, en el denominador, el flujo migratorio bajo hipótesis de

distribución uniforme de migraciones que se especifica en la ecuación como: $\frac{1}{2}N_{x-1}^{t} + \frac{1}{3}N_{x}^{t} + \frac{1}{6}N_{x}^{t+1}$. Por lo tanto, el segundo estimador queda; $\tilde{q}_x = \frac{\frac{1}{2}(D_x^t + D_x^{t+1})}{C_{x-1}^t - \frac{1}{2}D_{x-1}^t + \frac{1}{2}N_{x-1}^t + \frac{1}{3}N_x^t + \frac{1}{6}N_x^{t+1}}$. Para ambos estimadores, el número de defunciones se obtiene como la media entre las defunciones acaecidas en el año *t* y *t+1*, D_x^t y D_x^{t+1}, al estar basados en hipótesis de distribución uniforme de fallecidos para cada edad y año.

El tercer estimador modifica la naturaleza del estimador al considerar el momento exacto del fallecimiento de cada individuo que compone la población en un sistema demográfico cerrado (CDS_NH). De ese modo, se detraen el número de fallecidos en el año *t* y que nacieron en la cohorte *t-x*, denotado por $D_{x-1:t-x}^t$ de los individuos tenidos en cuenta al inicio del periodo C_{x-1}^t. Utilizando el mismo criterio, el numerador se obtiene tras tener en cuenta el número de defunciones ocurridas con edad *x* de los individuos nacidos en la misma cohorte: Los fallecidos con edad *x* durante el año *t* nacidos en la generación *t-x*, denotado por $D_{x:t-x}^t$ y los fallecidos con edad *x* durante el año *t+1* de la misma generación, *t-x*, denotado por $D_{x:t-x}^{t+1}$. Simplificando ambos términos, el estimador queda; $\breve{q}_x = \frac{D_{x:t-x}^{t,t+1}}{C_{x-1}^t - D_{x-1:t-x}^t}$.

Por último, el escenario que mayor información detallada requiere es aquel que incorpora el momento exacto de defunciones y migraciones en un sistema demográfico abierto (ODS_NH). Por lo tanto, al escenario descrito anteriormente debemos añadir el momento exacto de inmigración y el momento exacto de emigración que se recoge en la siguiente expresión: $E_{x:t-x}^{t,t+1} + \sum_{j=1}^{E_{x:t-x}^t} e_{x,j}^t + \sum_{j=1}^{E_{x:t-x}^{t+1}} e_{x,j}^{t+1} + \sum_{j=1}^{I_{x:t-x}^t} i_{x,j}^t + \sum_{j=1}^{I_{x:t-x}^{t+1}} i_{x,j}^{t+1}$. Así, el estimador AC libre de hipótesis implícitas quedará definido como sigue; $\ddot{q}_x = \frac{D_{x:t-x}^t + D_{x:t-x}^{t+1}}{C_{x-1}^t - D_{x-1:t-x}^t + N_{x-1:t-x}^t - E_{x:t-x}^{t,t+1} + \sum_{j=1}^{E_{x:t-x}^t} e_{x,j}^t + \sum_{j=1}^{E_{x:t-x}^{t+1}} e_{x,j}^{t+1} + \sum_{j=1}^{I_{x:t-x}^t} i_{x,j}^t + \sum_{j=1}^{I_{x:t-x}^{t+1}} i_{x,j}^{t+1}}$

De forma análoga se obtienen las correspondientes tasas de fallecimiento, m_x, para cada estimador. Este desarrollo, junto con un detalle exhaustivo del proceso comentado se puede encontrar en el trabajo (A2) del presente documento.

3.2. Estimador basado en el periodo

Para estimar las tasas de mortalidad en un estimador AP necesitamos conocer para cada edad y sexo el número total de defunciones y el tiempo total de exposición al riesgo de la población objeto de estudio durante el periodo de un año. La forma exacta en que se computa el tiempo total de exposición al riesgo dependerá del nivel de detalle disponible en los datos y, como se observa en el anexo estadístico (A3), en caso de asumir hipótesis implícitas, del razonamiento seguido.

Como punto de partida en este estimador, computamos el total de número de 'personas-año' de exposición al riesgo. Es importante remarcar que no partimos de entes demográficos teóricos que suelen ser representados en el esquema de Lexis (como el número total de personas que alcanzan la edad exacta *x* a lo largo del año *t*), sino que lo hacemos considerando el tipo datos que son habitualmente producidos por los sistemas estadísticos oficiales, i.e., stocks de poblaciones y flujos de migrantes y fallecidos. Así, bajo la hipótesis de distribución uniforme de fechas de nacimiento, obtenemos el número de personas expuestas al riesgo inicial como la media de la población registrada con edad completa *x* a 1 de enero del año *t*, C_x^t, y a 1 de enero del año *t+1*, C_x^{t+1}.

En los siguientes estimadores, ajustamos esta estimación inicial excluyendo/incluyendo el tiempo de exposición al riesgo de las personas que fallecen, inmigran o emigran con edad *x* durante el año *t,* incorporando también el momento exacto del nacimiento. A modo de resumen la Tabla 2 muestra los distintos estimadores dentro de la gama de un estimar AP para m_x.

Table 2. Resumen de los estimadores de un period-based estimator.

Estimador	Hipótesis			
	Sistema Demográfico Cerrado (CP)	Uniforme		
		Fallecidos (D)	Migraciones (M)	Fecha de Nacimiento (B)
$\widehat{m}_x = \frac{D_x^t}{\frac{1}{2}C_x^t - \frac{1}{6}D_x^t + \frac{1}{2}C_x^{t+1} + \frac{1}{6}D_x^t} = \frac{D_x^t}{\frac{1}{2}C_x^t + \frac{1}{2}C_x^{t+1}}$	Si	Si	N/A	Si
$\bar{m}_x = \frac{D_x^t}{\frac{1}{2}C_x^t - \frac{1}{6}D_x^t - \frac{1}{6}E_x^t + \frac{1}{6}I_x^t + \frac{1}{2}C_x^{t+1} + \frac{1}{6}D_x^t + \frac{1}{6}E_x^t - \frac{1}{6}I_x^t} = \frac{D_x^t}{\frac{1}{2}C_x^t + \frac{1}{2}C_x^{t+1}}$	No	Si	Si	Si
$\tilde{m}_x = \frac{D_x^t}{\frac{1}{2}C_x^t - \sum_{j=1}^{D_{x:U}^t}(1 - l_{x,j}^t) + \frac{1}{2}C_x^{t+1} + \sum_{j=1}^{D_{x:L}^t} d_{x,j}^t}$	Si	No	N/A	Si
$\ddot{m}_x = \frac{D_x^t}{\frac{1}{2}C_x^t - \sum_{j=1}^{D_{x:U}^t}(1 - l_{x,j}^t) - \sum_{j=1}^{E_{x:U}^t}(1 - l_{x,j}^t) + \sum_{j=1}^{I_{x:U}^t} i_{x,j}^t + \frac{1}{2}C_x^{t+1} + \sum_{j=1}^{D_{x:L}^t} d_{x,j}^t + \sum_{j=1}^{E_{x:L}^t} e_{x,j}^t - \sum_{j=1}^{I_{x:L}^t}(1 - l_{x,j}^t)}$	No	No	No	Si
$\dddot{m}_x = \frac{D_x^t}{\sum_{d=1}^{T}\frac{(d-0.5)}{T}C_{x,d}^t - \sum_{j=1}^{D_{x:U}^t}(1 - l_{x,j}^t) - \sum_{j=1}^{E_{x:U}^t}(1 - l_{x,j}^t) + \sum_{j=1}^{I_{x:U}^t} i_{x,j}^t + \sum_{d=1}^{T}\frac{(T-d+0.5)}{T}C_{x,d}^{t+1} + \sum_{j=1}^{D_{x:L}^t} d_{x,j}^t + \sum_{j=1}^{E_{x:L}^t} e_{x,j}^t - \sum_{j=1}^{I_{x:L}^t} I_{x,j}^t}$	No	No	No	No

N/A: No aplica.

El procedimiento llevado a cabo para incorporar cada una de las hipótesis en un estimador AP es el mismo que el desarrollado en un estimador AC. El estimador que menos información detallada requiere es aquel que considera sistema demográfico cerrado y distribución uniforme de defunciones y nacimientos (CP_UD_UB). Para este estimador la población expuesta al riesgo se obtiene como la media de la población al inicio y al final del periodo, C_x^t y C_x^{t+1}. En segundo lugar, y con distintas hipótesis, pero con el mismo output tras un amplio desarrollo estadístico (ver A3), tenemos el estimador que considera el sistema demográfico abierto y distribución uniforme de defunciones, migraciones y nacimientos (OP_UD_UM_UB). Para ambos estimadores la tasa de fallecimiento se obtiene utilizando la siguiente expresión: $\widehat{m}_x = \frac{D_x^t}{\frac{1}{2}C_x^t + \frac{1}{2}C_x^{t+1}}$.

El siguiente estimador se obtiene cuando se dispone del momento exacto de las defunciones, pero no se modifican el resto de hipótesis. En concreto, lo denotamos como sistema demográfico cerrado, no distribución uniforme de defunciones y distribución uniforme de nacimientos (CP_NUD_UB). De ese modo, el nuevo estimador se obtiene utilizando la siguiente expresión: $\tilde{m}_x = \frac{D_x^t}{\frac{1}{2}C_x^t - \sum_{j=1}^{D_{x:U}^t}(1 - l_{x,j}^t) + \frac{1}{2}C_x^{t+1} + \sum_{j=1}^{D_{x:L}^t} d_{x,j}^t}$.

Utilizando el mismo criterio, podemos desarrollar también el estimador cuando se tiene en cuenta el momento de la migración, que denotaremos como sistema demográfico abierto, no distribución uniforme de defunciones y migraciones distribución uniforme

de fechas de nacimientos (OP_NUD_NUM_UB), la expresión matemática resultante queda como sigue;

$$\ddot{\mathrm{m}}_{\mathrm{x}} = \frac{\mathrm{D}_{\mathrm{x}}^{\mathrm{t}}}{\frac{1}{2}C_x^t - \sum_{j=1}^{D_{x:U}^t}(1-l_{x,j}^t) - \sum_{j=1}^{E_{x:U}^t}(1-l_{x,j}^t) + \sum_{j=1}^{I_{x:U}^t} i_{x,j}^t + \frac{1}{2}C_x^{t+1} + \sum_{j=1}^{D_{x:L}^t} d_{x,j}^t + \sum_{j=1}^{E_{x:L}^t} e_{x,j}^t - \sum_{j=1}^{I_{x:L}^t}(1-l_{x,j}^t)}.$$

Por último, el nuevo estimador que se propone, incorpora, además del tiempo exacto de exposición al riesgo de fallecidos, inmigrantes y emigrantes, el tiempo exacto de exposición al riesgo de cada individuo que permanece con vida con edad *x* durante el año *t* como miembro de la población objeto de estudio. Así, una persona de edad *x* nacida el día *d* y que cumplió años durante el periodo *t* siendo, por ejemplo, d=1 el 1 de enero o d=32 el 1 de febrero, aporto $\frac{(d-0.5)}{T}$ años de exposición al riesgo, donde *T* es 365 o 366 para el año bisiesto. De igual modo, un individuo con edad *x* y que cumplió años durante el periodo *t-1*, nacido el día *d*, aportó $\frac{(T-d+0.5)}{T}$ años de exposición al riesgo. Así, incluyendo el resto de hipótesis, el estimador obtenido es el siguiente; $\breve{\mathrm{m}}_{\mathrm{x}} = \frac{\mathrm{D}_{\mathrm{x}}^{\mathrm{t}}}{\sum_{d=1}^{T}\frac{(d-0.5)}{T}C_{x,d}^t - \sum_{j=1}^{D_{x:U}^t}(1-l_{x,j}^t) - \sum_{j=1}^{E_{x:U}^t}(1-l_{x,j}^t) + \sum_{j=1}^{I_{x:U}^t} i_{x,j}^t + \sum_{d=1}^{T}\frac{(T-d+0.5)}{T}C_{x,d}^{t+1} + \sum_{j=1}^{D_{x:L}^t} d_{x,j}^t + \sum_{j=1}^{E_{x:L}^t} e_{x,j}^t - \sum_{j=1}^{I_{x:L}^t} l_{x,j}^t}$

En línea con el estimador AC, una explicación más detallada para la obtención de las probabilidades de fallecimiento, q_x, en un estimador AP se puede encontrar en el trabajo (A3) del presente documento.

4. Resultados

En el artículo (A1) las diferencias obtenidas entre las probabilidades brutas y graduadas de tener en cuenta o no los flujos migratorios son importantes. En concreto, utilizando un estadístico de disimilitud (Absolute Relative Error, ARE), observamos diferencias hasta del 4,5% para el rango de edades entre 14 y 36 años entre las probabilidades brutas y graduadas obtenidas con y sin incluir los flujos migratorios (ver Figura 1, A1). Concretamente, estas diferencias coinciden con la mayor intensidad en el saldo migratorio positivo acaecido en España durante el periodo 2006-2008. Por sexos, el impacto en el género femenino es menor que en el género masculino motivado por el mayor flujo migratorio en los hombres. Finalmente, los resultados de este trabajo alertan de la necesidad de analizar cada una de las hipótesis utilizadas en la tabla de mortalidad, así como su impacto en el sistema público de pensiones y en sector asegurador.

En esa línea, y dentro de los estimadores de tipo AC durante un periodo bianual, el trabajo (A2) profundiza el estudio llevado a cabo en el trabajo (A1) e incorpora, además del saldo migratorio (H2i), el momento exacto de la migración (inmigración y emigración), (H2ii), y el momento exacto del fallecimiento (H1). Por otro lado, también evalúa, utilizando una gran batería de test espaciales, paramétricos y funcionales, la aceptabilidad de ambas hipótesis.

Los resultados de los distintos test indican que la hipótesis de distribución uniforme de defunciones y migraciones y la hipótesis de sistema demográfico cerrado no son apropiadas. En concreto, tras observar los resultados de los contrastes espaciales no parece correcto asumir distribución uniforme de defunciones para el rango de edades entre 70 y 100 años, donde la intensidad de la mortalidad es mayor. Para la mayoría de edades comentadas se rechaza la hipótesis nula, H_0, de distribución uniforme de defunciones utilizadas en este estimador para todos los periodos analizados (ver Figuras 1A, 2A, 3, 3A y 4A, 2A). De igual modo, se evalúa si la distribución de los emigrantes e inmigrantes es uniforme para cada edad *x* y periodo analizado *t* (ver Figuras 4, 5, 5A, 6A, 7A, 8A, 9A, 10A, 11A y 12A, A2). De nuevo, se observa un rechazo generalizado en los test de hipótesis en las edades con mayor intensidad en el flujo migratorio. Por otro lado, una serie de contrastes paramétricos (ver Figuras 6, 7 y 8) han

evaluado los tiempos medios que han estado expuestos al riesgo para cada edad *x* y periodo de *t* cada evento demográfico. Por ejemplo, se ha evaluado si por término medio el tiempo vivido por los fallecidos y emigrantes (inmigrantes) ha sido ½, con resultados muy similares a los observados en los contrastes espaciales. Finalmente, dentro de la hipótesis de sistema demográfico cerrado (H2), se evalúa si el número de emigrantes (inmigrantes) es significativo dado un nivel de probabilidad del 0,5% y si el número de inmigrantes es similar al número de emigrantes (ver Figuras 9 y 10).

El rechazo observado en la gran batería de contrastes realizados tiene implicaciones directas en las probabilidades de fallecimiento, q_x. De nuevo, utilizando un estimador de disimilitud (Absolute Relative Discrepancy, ARD) el trabajo A2 muestra diferencias en torno al 4% entre el estimador que tiene en cuenta el sistema demográfico cerrado con no hipótesis de distribución de los fallecidos y el sistema demográfico abierto con no hipótesis de distribución de los fallecidos (ver Figura 11, A2). A pesar que los contrastes realizados rechazan, en las edades avanzadas, la distribución uniforme de defunciones, su impacto en las estimaciones de las probabilidades brutas y graduadas de fallecimiento son mínimas (ver Figuras 44A, 49A y 51A, Supplementary material A2), a lo sumo un 1,50%.

Para analizar las diferencias en las probabilidades de fallecimiento en los seguros de vida utilizamos un producto de seguros destinado a cubrir la contingencia de supervivencia (renta vitalicia, *annuity*) y un producto destinado a cubrir la contingencia de fallecimiento (temporal anual de riesgo, *premium*). La tabla 4 (artículo A2) muestra las diferencias (en días para la renta y en porcentaje para el temporal anual de riesgo) al comparar cada uno de los estimadores utilizados. Además, en la comparación se incluyen las estimaciones realizadas por los agentes oficiales del INE y el Human Mortality Database (HMD). Como se observa en dicho cuadro, en el producto de rentas vitalicias las diferencias más grandes se dan entre el estimador que tiene en cuenta todas las hipótesis (sistema demográfico cerrado y distribución uniforme de nacimientos y migraciones) y el estimador que no tiene en cuenta ninguna de las hipótesis implícitas (sistema demográfico abierto y no hipótesis de distribución uniforme de nacimientos y migraciones). En concreto, hasta 11 (15) días de diferencia al considerar una renta vitalicia, $\ddot{a}_x$ para los hombres (mujeres). Las mayores diferencias, hasta de 153 (48) días

en hombres (mujeres), se observan cuando se comparan las estimaciones desarrollados en el trabajo con las estimaciones obtenidas por los agentes económicos oficiales (INE o HMD). Las diferencias en los productos que cubren la contingencia de fallecimiento son también importantes. El precio a pagar por una persona de 20 años para cubrir la contingencia de fallecimiento durante un año, ${}_1\mathrm{A}_{20}$, es, hasta un 8,33% (4,69%) en hombres (mujeres), más barato cuando se introducen los flujos migratorios en las estimaciones de probabilidad. Diferencias como las observadas en este trabajo, permitirían la reducción (incremento) en los precios de los seguros de vida en un saldo migratorio positivo (negativo) como el observado (ver figuras 44-S, A3) hasta el año 2008 (a partir del año 2008).

En la actualidad, el estimador AP es el más utilizado por los organismos oficiales para estudiar la mortalidad de una población (ver, por ejemplo, ONS 2010, 2012, INE 2009, 2016 y Arias 2015). En el artículo (A3) se desarrollan nuevos estimadores que, además de hacer innecesarias las dos primeras hipótesis (H1) y (H2) descritas en el trabajo (A2), prescinden de la tercera hipótesis (H3): el desconocimiento del momento exacto del nacimiento (cumpleaños) de cada uno de los individuos que componen la población.

Adicionalmente, en el trabajo (A3) se demuestra matemáticamente que el desarrollo teórico para obtener el número de expuestos al riesgo depende del punto de origen de los datos, lo que genera estimadores diferentes cuando son utilizadas las hipótesis simplificadoras (H1, H2 y H3). Por ello, una gran fortaleza de este trabajo es el desarrollo teórico (Anexo, A3) que permite obtener un estimador libre de inconsistencias teóricas como las encontradas en los estimadores utilizados actualmente por los agentes oficiales. Por ejemplo, Wilmoth *et al.* (2007) parten del número de personas que alcanzan la edad exacta *x* y la edad exacta *x+1* a lo largo del año *t* (líneas AD y BC Figura 1, A3), cuando en la práctica, lo habitual es disponer del número de personas que se encuentran con vida con edad *x* en un momento concreto del tiempo, habitualmente 1 de enero del año *t* y *t+1* (líneas AB y CD Figura 1, A3).

Los resultados empíricos del trabajo A3 muestran (Figura 7, A3) que, para la base de datos estudiada, las mayores discrepancias cuando se tiene en cuenta el momento del nacimiento son atribuibles a las personas nacidas durante la Guerra Civil española,

especialmente al final de la contienda. El mayor número de nacimientos localizados en pocos meses provoca diferencias en las estimaciones de tasas de mortalidad de hasta un 4% y 5% en los hombres y mujeres respectivamente. La no aceptabilidad de la hipótesis de distribución uniforme de fechas de nacimiento (cumpleaños) tiene, como en los otros trabajos, repercusión en productos de seguros y en el sistema de pensiones.

Cuando se compara el estimador que tiene en cuenta el sistema demográfico abierto con no hipótesis sobre los fallecidos y migraciones e hipótesis de distribución uniforme de cumpleaños (OP_NUD_NUM_UB) con el estimador libre de hipótesis, sistema demográfico abierto con no hipótesis sobre los fallecidos, migraciones y cumpleaños (OP_NUD_NUM_NUB) las diferencias en las rentas son reducidas (ver Tabla 3, A3). Hasta un 0,26% y 0,12% en hombres y mujeres respectivamente en una renta actuarial prepagable temporal de 30 años para la edad de 35 en las personas que nacieron durante la Guerra Civil, ${}_{30}\ddot{a}_{35}^{1940}$. Sin embargo, hay un dato muy relevante en la tabla comentada. Mientras el estimador OP_NUD_NUM_UB no recoge el efecto en los mayores nacimientos una vez finalizada la contienda, ${}_{10}\ddot{a}_{55}^{1941} = 871{,}31\ €$ frente a ${}_{10}\ddot{a}_{55}^{1940} = 874{,}79\ €$, el estimador OP_NUD_NUM_NUB sí recoge dicho efecto en el valor de la renta temporal, ${}_{10}\ddot{a}_{55}^{1941} = 872{,}86\ €$ frente a ${}_{10}\ddot{a}_{55}^{1940} = 872{,}81\ €$.

Las diferencias en los productos que cubren la cobertura de fallecimiento (Premium) son más notables. Por ejemplo, el precio al contratar un seguro temporal anual de riesgo renovable que cubra la contingencia de fallecimiento de importe 100.000€ con 60 años de edad es un 1,88% (1,95%) más barato en hombres (mujeres) usando el estimador OP_NUD_NUM_NUB que el estimador OP_NUD_NUM_UB. Especial atención recibe la generación nacida durante la Guerra Civil. Las diferencias son hasta un 4,68% para hombres y un 5,05% para mujeres al comparar ambos estimadores de los individuos nacidos en 1939 o 1940. De nuevo, las diferencias son sumamente elevadas para un producto ampliamente comercializado en el sector asegurador.

La no aceptabilidad de la hipótesis de distribución uniforme de nacimientos pone en alerta posibles repercusiones en otras disciplinas de las ciencias sociales. Así, en el trabajo (A4) se analizada la distribución anual de nacimientos a lo largo de todo el siglo XX. Concretamente este trabajo ahonda en la literatura sobre la estacionalidad de los nacimientos al estudiar su distribución y evolución durante las últimas décadas.

Utilizando datos del padrón de habitantes de la Comunitat Valenciana y aplicando técnicas estadísticas como las propias de series temporales, la investigación (A4) aporta evidencia empírica de cómo la organización actual de los tiempos de trabajo en el sector sanitario está impactando sobre la distribución semanal de nacimientos, al imponer el cuerpo médico su posición hegemónica. Se constata un cambio en la distribución de nacimientos, por días de la semana, desde mediados del siglo XX hasta la actualidad.

5. Conclusiones

El desarrollo llevado a cabo en esta tesis doctoral permite disponer de una gama de estimadores libre de hipótesis implícitas, dentro de la familia del modelo de dos factores, para la construcción de tablas de mortalidad. De manera paralela, el trabajo ha evaluado las distintas hipótesis empleadas históricamente en la construcción de estas tablas de mortalidad. Además, la investigación ha ampliado el estudio de la estacionalidad de los nacimientos desde una nueva perspectiva en el campo de las ciencias sociales: el estudio de la estacionalidad semanal de los nacimientos. En consecuencia, múltiples conclusiones son obtenidas tras la redacción de este trabajo.

Tras analizar los resultados obtenidos en la presente investigación se concluye que la asunción de determinadas hipótesis en las estimaciones de probabilidad y tasa de fallecimiento no son adecuadas. En primer lugar, podemos concluir que asumir distribución uniforme de fallecidos no es correcto para un gran rango de edades. Además de la distribución no uniforme de los fallecidos ya conocida a la edad de los 0 años, con gran concentración de fallecidos en los primeros días/semanas de vida, la distribución de los fallecidos en las edades adultas tampoco es uniforme al observar el gran número de rechazos en los distintos test de hipótesis.

En la actualidad, la comercialización de productos destinados a cubrir la contingencia de fallecimiento (productos de vida-riesgo que aseguran un capital en caso de fallecimiento, A_x) se está ampliando hasta los 75 años como consecuencia del incremento continuado de la esperanza de vida. De igual modo sucede en los productos de rentas vitalicias utilizados en el sistema público de pensiones y el sector asegurador. Una diferencia en las probabilidades de fallecimiento en las edades entre los 60 y 75 años puede ser un problema en las estimaciones económicas realizadas por los agentes económicos. Por ese motivo, se concluye que los resultados de este trabajo motivan la necesidad de realizar un análisis pormenorizado en las tablas de mortalidad utilizas por los organismos públicos y privados.

Por otro lado, tras los resultados observados en los distintos contrastes realizados podemos concluir que la hipótesis de sistema demográfico cerrado no es adecuada, al menos para las dos poblaciones estudiadas en este trabajo. El flujo migratorio positivo (negativo) acaecido hasta el año 2008 (a partir del año 2008) afecta

a las estimaciones de probabilidad en un estimador AC. Las diferencias observadas, hasta del 4% en las edades de mayor flujo migratorio, nos permite concluir que es necesario eliminar dicha hipótesis en un país (región) que se caracteriza por una intensidad elevada de sus flujos migratorios. Se conoce que el signo del flujo migratorio esté condicionado a la situación económica de la zona de estudio. Así, en épocas de expansión económica, será esperable que el flujo de inmigrantes sea mayor que el flujo de emigrantes. Se concluye que este acontecimiento puede incrementar (decrementar) la población expuesta al riesgo con la correspondiente disminución (incremento) de la probabilidad de fallecimiento.

La distribución de nacimientos ha sido la última hipótesis analizada y evaluada en el presente trabajo. Podemos concluir que los distintos eventos históricos acontecidos en nuestro país y los comportamientos socio-culturales y de organización evidencian que no es adecuado asumir tampoco la distribución uniforme de nacimientos. El escaso número de nacimientos acontecidos durante la Guerra Civil y su repunte una vez terminada la contienda es un claro ejemplo. Por otro lado, la estacionalidad de los nacimientos se ha estudiado durante años en la literatura sociológica. Cada vez son más, y sobre la base de evidencias empíricas y utilizando modelos estadísticos, los estudios que sitúan los comportamientos sociales y culturales, con predominio sobre los estudios de fotoperiodo, como factor explicativo del comportamiento de los nacimientos en una región. El trabajo (A4) ahonda en la literatura de la estacionalidad de los nacimientos desde una nueva perspectiva, muy poco estudiada hasta el momento: el análisis en la distribución de los nacimientos durante la semana. Este estudio, junto con los múltiples trabajos citados en el mismo, evidencian que cada vez más las formas de organización del tiempo en una sociedad junto con los cambios socio-culturales están impactando en la distribución de los nacimientos. Se concluye que estas prácticas de organización social pueden tener un impacto considerable en la hipótesis de distribución uniforme de nacimientos.

El desarrollo analítico ha permitido obtener nuevos estimadores para el modelo de edad-cohorte y para el modelo de edad-periodo libres de hipótesis implícitas. Cabe recordar que el estimador desarrollado de edad-periodo está libre de las inconsistencias teóricas observables con los estimadores actuales. A la luz de los resultados obtenidos

en este trabajo, abogamos por la utilización de los estimadores propuestos en futuros desarrollos llevados a cabo tanto por el INE para estimaciones públicas como por la DGS para el sector asegurador.

Tras la explosión de información disponible actualmente y el desarrollo de los nuevos sistemas informáticos ha quedado demostrado que el coste de introducir el momento exacto de fallecimientos, flujo migratorio y nacimientos es mínimo. Sin embargo, el potencial impacto en las probabilidades y tasas de mortalidad utilizadas por organismos públicos y privados es elevado.

6. Futuras líneas de investigación

Es preciso notar que, aunque la tesis doctoral se centra en analizar el impacto de las tres hipótesis (H1), (H2) y (H3) mencionadas, la construcción de las tablas de mortalidad incluye la asunción de dos hipótesis implícitas adicionales (H4) y (H5), cuyo análisis es cualitativamente mucho más complejo. En concreto, tanto en el estimador basado en el periodo (AP) como en el estimador basado en la cohorte (AC) asumen implícitamente que los inmigrantes (emigrantes) adquieren (poseen) el mismo riesgo de fallecimiento que la población residente equivalente y, que, por tanto, en la decisión migratoria no opera ningún efecto de selección. Por lo tanto, el riesgo de muerte es independiente, a cada edad, de la historia personal del individuo que ha migrado. Esta hipótesis no parece razonable en un escenario con flujos migratorios, por lo que una posible línea futura de investigación sería el estudio de la misma. Su evaluación y análisis, sin embargo, es sumamente complejo. Para ello, se necesitaría disponer de datos longitudinales, información que, al menos para el caso de España, no está disponible.

En este trabajo se ha realizado un análisis profundo de la obtención de cada uno de los dos estimadores, de tipo AP y AC, del modelo de dos factores. Sin embargo, quedaría por comparar los resultados de ambos estimadores sobre una misma base de datos. Es de esperar una diferencia entre ellos pues, aunque ambos utilicen las dos primeras hipótesis (H1) y (H2) no comparten la hipótesis (H3) y el procedimiento de su obtención difiere.

El estimador AC desarrollado en este trabajo analiza la mortalidad para un periodo bianual. Sin embargo, una nueva línea de investigación abarcaría todo el desarrollo teórico y empírico de un nuevo estimador en los países como Reino Unido o Australia que utilizan tres años en sus estimaciones de probabilidad. Para ambos países sería de gran importancia evaluar de nuevo, mediante contrastes de hipótesis, cada una de las hipótesis H1 y H2. Los movimientos demográficos acaecidos en los países comentados difieren de los observados en España, motivo por el cual nuevos y valiosos resultados pueden ser obtenidos.

En el marco del estimador AP la no aceptación de la hipótesis de distribución uniforme de nacimientos abre una nueva línea de investigación en el marco

internacional. La participación directa de algunos países como Reino Unido, Alemania o Francia en las dos guerras mundiales, especialmente la segunda, al ser más reciente, puede mostrar, al igual que sucede con la Guerra Civil en España, que dicha hipótesis tenga grandes efectos para las cohortes nacidas en esos años. Los países mencionados poseen una cultura aseguradora desarrollada, donde gran parte de la población destina una parte de sus ingresos a productos de seguros de vida para cubrir la reducción de poder adquisitivo durante la etapa pasiva de jubilación. Sería de enorme interés estadístico-actuarial disponer de un estimador que tuviera en cuenta los menores (mayores) nacimientos concentrados en periodos cortos de tiempo durante (después de) los eventos comentados.

Por otro lado, en la valoración de los productos de seguros destinados a cubrir la contingencia de supervivencia es habitual utilizar una tabla de mortalidad generacional que recoge la experiencia de la mortalidad de cada cohorte completa. La construcción de esta tipología de tabla de mortalidad es sumamente compleja debido a la información necesaria. Actualmente, para su construcción se utilizada información histórica agregada correspondientes a las defunciones y la población expuesta al riesgo para cada edad y año. Información que actualmente está disponible en el Human Mortality Database (HMD). De nuevo, quedaría por desarrollar la metodología para construir una tabla de mortalidad generacional libre de las hipótesis (H1), (H2) y (H3). Para su obtención, se necesita disponer de información detallada, especialmente para construir un estimador AP, de cada evento demográfico para una gran cantidad de años y edades. En la actualidad no se dispone de información de micro-datos a tal nivel de detalle. Quizá, una primera aproximación en la hipótesis (H3) sería utilizar la distribución de los nacimientos de las personas con vida en una fecha concreta. De igual modo, la hipótesis (H2) podría ser mitigada ante el conocimiento de la distribución de los migrantes con vida también en una misma fecha.

Otra posible línea de investigación en estadística e investigación operativa sería el desarrollo de un *package* en R que implementara todos los estimadores desarrollados en esta investigación. Disponer de este *package* permitiría abordar de una manera más rápida y eficiente las distintas investigaciones comentadas anteriormente.

En el marco del sector asegurador se estudia *la mortalidad base* y la *tendencia de la mortalidad*. Por un lado, el primer concepto analiza la mortalidad en un momento concreto del tiempo. Para su estudio lo habitual es utilizar uno de los dos estimadores desarrollados en el presente trabajo. Por otro lado, y con cada vez mayor énfasis en la literatura científica, se analiza la mortalidad futura o *tendencia de la mortalidad* para analizar el riesgo de longevidad. Los numerosos trabajos en este ámbito (ver Lee y Carter 1992, Currie *et al.* 2004, Li y Lee 2005, Cairns *et al.* 2008, 2009, Haberman y Renshaw 2011, Russolillo *et al.* 2011, Danesi *et al.* 2015, Cains *et al*. 2016 y Enchev *et al.* 2016) utilizan la tasa de mortalidad sin incluir las hipótesis desarrolladas en la presente tesis doctoral. Lo habitual es utilizar la población al inicio o al final de cada año e incluir el número de fallecidos bajo hipótesis de distribución uniforme. Podría ser de gran interés en la literatura analizar la tendencia de la mortalidad excluyendo las hipótesis implícitas mencionadas. En la actualidad, desconocemos si el mayor número de nacimientos acontecidos una vez terminada la Guerra Civil, o la no distribución uniforme de nacimientos (A3) y la eliminación de la hipótesis de sistema demográfico cerrado (A2) tiene alguna implicación en la tendencia de la mortalidad futura.

Finalmente, dentro de esta línea de investigación quedaría también por analizar la derivación de las tablas de mortalidad en el sector asegurador. Generalmente, las tablas de mortalidad son utilizadas en el sector asegurador para procesos como el cálculo de reservas (*reserving*) y el de tarificación o de precios (*pricing*). Es habitual que estas tablas de mortalidad se construyan con datos agregados e incorporen recargos de seguridad como medida de prudencia. Desde la entrada de la nueva normativa reguladora de Solvencia II, el pasado 1 de enero de 2016, las compañías aseguradoras deben calcular las provisiones técnicas utilizando hipótesis libres de riesgo (best-estimate) como mecanismo de trasparencia. Sin embargo, actualmente las compañías de seguros aplican un porcentaje de la tabla de mortalidad para aislar el componente libre de riesgo y el margen o recargo de seguridad. Este sencillo procedimiento conlleva la asunción de algunas limitaciones como por ejemplo que la tabla libre de riesgo de la compañía de seguros tiene el mismo comportamiento edad a edad que la tabla de mortalidad general. Sería de suma importancia construir una tabla de mortalidad a partir de la propia experiencia de la compañía aseguradora (combinándola si es preciso con la

de la población general utilizando aproximaciones propias de la teoría de la credibilidad y/o de la estadística bayesiana). Haciendo uso de un estimador basado en la cohorte o periodo (extensión de un estimador AC o AP), se podría separar el componente libre de riesgo y el margen de seguridad, mejorando los resultados que se obtienen con las metodologías más habituales.

Introducing migratory flows in life table construction (A1)

Jose M. Pavía, Universitat de Valencia

Francisco G. Morillas, Universitat de Valencia

Josep Lledó, Universitat de Valencia

Abstract

The purpose of life tables is to describe the mortality behavior of particular groups. The construction of general life tables is based on death statistics and census figures of resident populations under the hypothesis of closed demographic system. Among other assumptions, this hypothesis implicitly assumes that entries (immigrants) and exits (emigrants) of the population are usually not significant (being almost of the same magnitude for each age compensating each other). This paper theoretically extends the classical solution to open demographic systems and studies the impact of this hypothesis in constructing a life table. In particular, using the data of residential variations made available to the public by the Spanish National Statistical Office (INE, Instituto Nacional de Estadística) to approximate migratory flows, we introduce in the process of constructing a life table these flows and compare, before and after graduation, the crude mortality rates and the adjusted death probabilities obtained when migratory flows are, and are not, taken into account.

Keywords: General population, Lexis diagram, open demographic system.

Acknowledgements

The authors wish to thank M. Hodkinson for translating into English the text of the paper and the anonymous referees their valuable comments and suggestions. The authors acknowledge the support of the Spanish Ministry of Science and Innovation (MICINN) through the project CSO2009-11246.

1. Introduction

In the demographic and actuarial fields, the analysis of mortality in a population has particular relevance for their applications. Life tables, or mortality tables, are used to recreate an observed mortality situation or to present future values of the evolution of mortality in certain groups, making it possible to generate demographic forecasts or to calculate premiums and/or income for life insurance and pension benefits. Medicine is another area where mortality analysis is also frequently used.

Life tables are usually drawn from the study and analysis of the intensity and rate at which mortality affects each age group in question. In general populations, it is generated using information mainly from population censuses and lists of deceased, where individual records from insurance policies are the prime source of information in insured populations. To be specific, and once the sample period has been decided, the comparison between the numbers at risk and the number of deaths allows the actuary (demographer) to obtain initial (crude) estimates for the probability of death in each age group q_x. These probabilities are subjected to the corresponding graduation or adjustment processes (see, e.g., Copas-Haberman, 1983; Forfar *et al.*, 1988; or Ayuso *et al.*, 2007) with a view to smoothing the profile of the associated stochastic process and to ultimately develop appropriate tables from a fictitious starting population of size l_0.

In the construction of mortality tables for general populations (which is the subject of this paper) it is not usual to specifically consider migratory flows, making the hypothesis of closed demographic system (HCDS), which implicitly entails the assumption of certain limitations, the main ones being: (i) that migration flows (inputs and outputs) of the population by age and sex are considered not to be significant; (ii) that for each age group migration flows are random and show similar entry and exit figures; and, (iii) that immigrants acquire the same risk of death as the resident population.

These limitations, which are not always reasonable, should be checked because of their potential impact on, for example, the calculation of life expectancy, of premiums for life insurance or of estimates for the calculation of pensions. The aim of this paper is twofold, firstly, to introduce an estimator for an open demographic system and, secondly, to show the incidence of HCDS through a real case. More specifically, given

the immense pressure of migration endured by Europe, and in particular Spain, in recent years, the analysis will be based on the comparison of mortality tables (by gender) obtained for Spain under HCDS and under the hypothesis of open demographic system (HODS), in which migratory flows are explicitly considered. The comparison is carried out in two ways. Firstly, the differences between the estimated crude probabilities obtained under each hypothesis are compared and, secondly, the comparison is again carried out after having graduated the crude data.

The rest of the paper is structured as follows. Section 2 explains the methodology used to obtain the mortality tables, both for the closed demographic system and the open demographic system. In Section 3, different comparisons are carried out between the HCDS and HODS tables. The last section presents the conclusions reached and indicates several issues for future research.

2. Methodology

The techniques and formulae used to estimate life tables are heavily influenced by the type of information available. When working with official statistics, the relevant data are usually offered, by and large, in aggregate form. Hence, in this research we have opted to work with aggregated figures –which come from information that the Spanish National Statistical Office (INE) has made available on its website (http://www.ine.es) – despite it being possible in the Spanish case to use some detailed (anonymized) microdata[4]. In particular, we will consider that, for each gender, aggregated figures of migrants, deaths and population are available by age and calendar year. Under these circumstances, the representation and analysis of the information in a Lexis diagram (named after the German statistician, economist and social scientist Wihelm Lexis, who adopted it in the nineteenth century to illustrate the procedures for calculating a mortality table) often greatly facilitates the reasoning and makes more manageable,

[4] Such is the case of death statistics and of residential variation figures (directly available on the INE website). Indeed, the INE, unlike many other national and international statistical agencies, is characterized by its willingness to make available, in anonymized form, the detailed data generating the vast majority of its statistical operations.

after the application of a number of reasonable hypotheses, handling the flaws of detailed information which are present in aggregate data[5].

The Lexis diagram is a two-dimensional diagram of lifelines with two-temporal dimension: calendar time and age. Calendar time is represented on the horizontal axis, while age is represented on the vertical axis. This diagram permits representation of the life events of a population from the personal history of the individuals within it. Each personal story is represented by a line segment forming an angle of 45 degrees to the horizontal axis. The classical approach (of closed demographic system) states that each personal story begins at birth, which is represented on the baseline, and ends at some point on the graph with the individual's death (Livi Bacci, 2000). In this paper, however, personal history is not determined solely by the events of birth and death. The introduction of migratory flows makes it necessary to modify the classical interpretation of the Lexis diagram since, in this case, the history of an individual could start from birth or from immigration and, likewise, it could end by death or by emigration.

Figure 1 shows a small section of a Lexis diagram which, as usual, comes divided into cells of dimensions of 1×1 so that between each pair of oblique lines are the lifelines that make up a generation of individuals and where each cell represents an observation period of one year (in which the age of the individuals also has a variation of a year).

For example, in Figure 1 (left) various individual life lines (thin lines) have been represented. Lifelines of individuals entering in the system due to immigration can originate anywhere in the diagram and are differentiated from the rest by having a ○ at its origin. When an individual leaves the system, the cause of departure is differentiated graphically: if the reason is death, the tail end of the lifeline is marked with an x; if the reason is emigration the end-lifeline is marked with a □.

[5] Obviously, some of the hypotheses to be considered could be relaxed using more detailed data (see, e.g., INE, 2009).

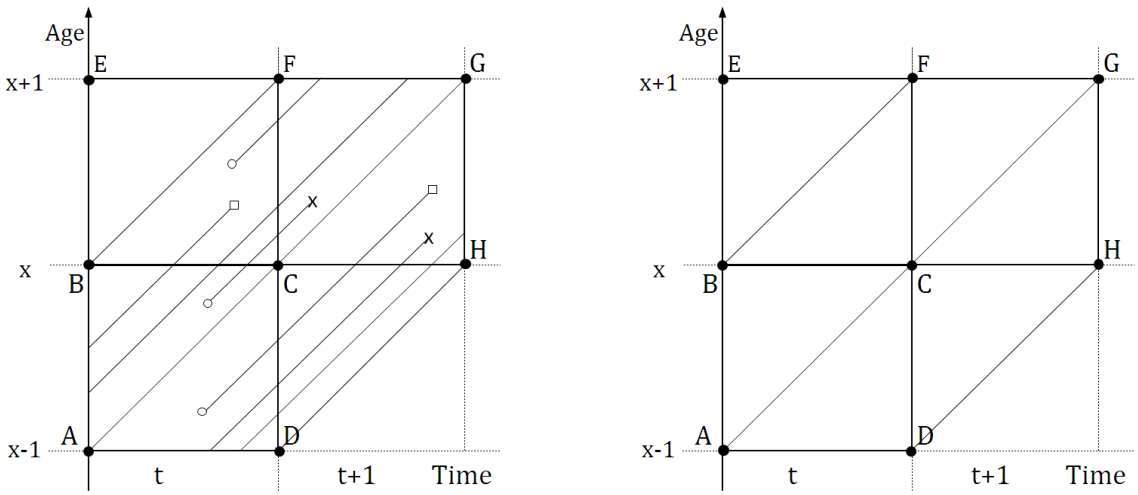

Figure 1: Detail (2×2) of Lexis diagram with lifelines (left) and schematic (right).

The aggregate information, however, does not allow for an accurate location of the lifelines of the individuals who comprise the study population. To make use of this geometric representation, therefore, the usual convention of the Lexis scheme of assigning values to segments and surfaces must be adopted. In this paper, the value of a segment is always identified by the number of lifelines that cross it, while the value(s) to be associated with each area will depend on the hypothesis under consideration.

2.1. Closed demographic system

Under HSDC, each surface is identified with a single value: the number of lifelines that end in it due to death. So, assuming for simplicity that (as in our case) for any given age *x* there are available the number of residents counted in January 1 of year *t*, C_x^t, and the number of residents who died in each age *x* for years *t* and *t+1* (D_x^t and D_x^{t+1} respectively), we have that, as shown in Figure 1 (right), it is straightforward to draw the information. To be specific, on the one hand, the quadrilaterals ABCD, BCEF and CFGH will be identifiable respectively with D_{x-1}^t, D_x^t and D_x^{t+1}, and, on the other hand, the segment AB will be equal to C_{x-1}^t.

At this point, it is now easy to obtain an initial estimate of q_x exploiting the geometric properties of the representation and assuming uniform distribution of birth dates and deaths within each age group and year. In particular, noting that the segment BC represents the number of people reaching age *x* in year *t*, it follows that under HSDC

the number of them who die before reaching age *x+1* will come represented by the quadrilateral BCFG and that therefore an estimate for q_x is obtained from[6]:

$$\widehat{q_x} = \frac{BCFG}{BC} \tag{1}$$

And from this, using the geometry of the scheme, one arrives at $BCFG = BCF + CFG$ and $BC = AB - ABC$; from which, using the hypothesis of uniform distribution, one deduces $BCFG = \frac{BCEF}{2} + \frac{CFGH}{2}$ and hence:[7,8,9]

$$\widehat{q_x} = \frac{\frac{1}{2}(D_x^t + D_x^{t+1})}{C_{x-1}^t - \frac{1}{2}D_{x-1}^t} \tag{2}$$

2.2. Introducing migratory flows

Assuming HODS, the expression (2) becomes invalid and another approach is required to take into account the entries and exits that happen in the study group during the analysis period. At this point, it will be useful to refer to the type of reasoning usually employed for insured groups (see, e.g., Benjamin and Pollard, 1992), where the number

---

[6] Given that $q_x = \frac{d_x}{l_x}$ is defined as the quotient of the number of deaths between ages *x* and *x*+1, *dx*, and the number of survivors at age *x*, l_x.

[7] This general expression, nevertheless, would not be appropriate for ages zero and one, since as it is well-known the deaths of children under one year old are concentrated in their first weeks of life. The assumption of uniform distribution cannot therefore be maintained for deaths counted with zero years: the greater part of these deaths will be located in the corresponding lower triangle. Thus, the formulae used for ages zero and one have been, respectively $\widehat{q_0} = \frac{0.7D_0^t + 0.3D_0^{t+1}}{B^t}$ (where *Bt* denotes the births in year *t*) and $\widehat{q_1} = \frac{\frac{1}{2}(D_1^t + D_1^{t+1})}{C_0^t - 0.3D_0^t}$; which can be obtained assuming that the number of deaths occurring during the first half of age zero is approximately four times the number of deaths registered during the second half. Obviously, if the deaths by generation (also available on the INE website; INE, 2010a) were used, no hypothesis about how to distribute the deaths between the triangles would be necessary because of the values of these would be known exactly.

[8] Unlike the expression used to estimate q_x in this paper, until recently the INE started with $BC = CF + BCF$ and arrived at a different equation $\widehat{q_x} = \frac{\frac{1}{2}(D_x^t + D_x^{t+1})}{C_x^{t+1} + \frac{1}{2}D_x^t}$ although equivalent under HCDS (INE, 2007). Since 2009, the INE estimates central age-specific death rates, mx, using the detailed information available on death microdata to obtain in each age the exact time spent by those who die during the year of study (see, INE, 2009). The use of those detailed data makes it unnecessary to assume any hypotheses about the distribution of deaths within each age and calendar year.

[9] A general formula to estimate q_x when the census of the population is located at any instant of the year and not necessarily at the start can be found in, for example, Pavía (2011, Ex. 91).

of deaths observed is separated depending on the risks of death and the time exposed to risk. That is, under HODS, the number of deaths observed, BCFG, will be approximately equal to the number of people that reach age *x*, BC, by the probability of any of them dying before reaching the age *x+1*, q_x, plus the number of people that immigrate with age *x+k* (where 0 < k < 1), whose lifeline starting point would be located in the surface BCFG, by the probability that a person of age *x+k* dies before reaching age x+1, ${}_{1-k}q_x$[10] minus the number of people that emigrate with age *x+k* (where 0 < k < 1), whose lifeline end point would be located in surface BCFG, by the probability that a person of age *x+k* dies before reaching the age *x+1*, ${}_{1-\mathrm{k}}\mathrm{q}_\mathrm{x}$.

The problem is that, unlike what happens in insured populations, the dates and specific ages at which a person immigrates or emigrates are not usually known, so to obtain a useful expression of the decomposition of BCFG it is essential to extend the classic convention and assign additional variables to each surface of the Lexis diagram. To be specific, we propose to link to each area two new variables: the number of lifelines that start in the surface (immigrants) and the number of lifelines that finish in the surface for reasons other than death (emigrants). With this extension it will then be possible to obtain an operative expression for BCFG from which an estimator for q_x can be derived by simply adding hypotheses (i) on the distribution of entries and exits in each surface (which, in the same way as death distribution, are assumed to be uniform, since it is reasonable for the level of information available) and (ii) on the risk of death throughout each age *x* (which as a rule is assumed proportional to the period of exposure to risk, that is, ${}_{1-\mathrm{k}}\mathrm{q}_\mathrm{x} = (1-\mathrm{k})\mathrm{q}_\mathrm{x}$).

As usual when handling official statistics, it is assumed that the total number of people that immigrate and emigrate in any year *t* and with age *x* is only known in aggregate terms, I_x^t and E_x^t, respectively. Obviously, other more informative situations in which more precise data about the age distribution of migrants in each year were available, for example from microdata of residential variations, would also be perfectly treated with this strategy.

[10] Note that using this expression entails the implicit assumption that immigrants acquire the same risk of death as the population in which they integrate.

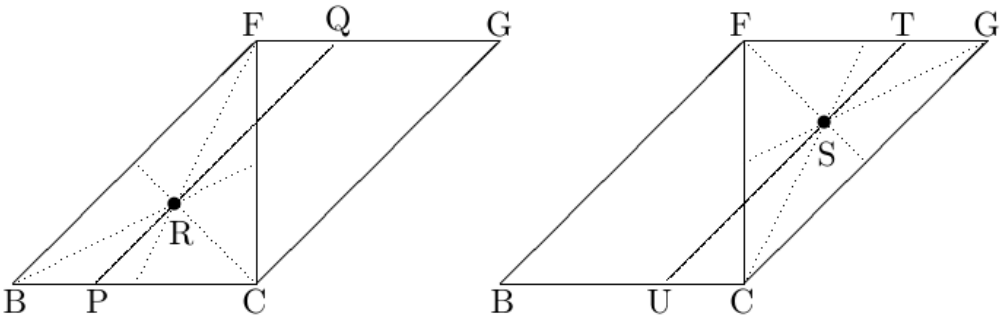

Figure 2: Barycentres of surfaces of migratory movements.

Under the conditions above, denoting by $N_x^t = I_x^t - E_x^t$ the net migration recorded in year *t* at age *x*, it follows that the number of people who reach age *x* during year *t*, BC, would be equal to $BC = C_{x-1}^t - \frac{1}{2}D_{x-1}^t + \frac{1}{2}N_{x-1}^t$, and that the number of entries and exits that would be registered in each of the triangles BCF and CFG would be, respectively, $\frac{I_x^t}{2}$, $\frac{E_x^t}{2}$, and $\frac{I_x^{t+1}}{2}$, $\frac{E_x^{t+1}}{2}$, and likewise, under the same hypotheses, each exit/entry produced in each triangle would be located, in average terms, in the centroid (barycentre) of the corresponding triangle (see Figure 2).[11]

As can be seen in the representation on the left of Figure 2, the point R is in the barycentre or centroid of the triangle BCF, which is easy to prove to be at a distance of $\frac{2\sqrt{2}}{3}$ from point Q, in the same way that point S (which can be taken as representative of all points in which an entry/exit occurs in triangle CFG) is at a distance $\frac{\sqrt{2}}{3}$ from T.[12] From

[11] Alternatively, it can be demonstrated that the average of the distances (across the lifelines) of each point of the corresponding triangle to segment *FG* is equal to the distance of the corresponding barycentre to segment *FG*. For example, considering the triangle *BCF* and, inside it, an arbitrary point *K* with coordinates (*x*,*y*), it is not difficult to prove that the lifeline of *K* intersects *FG* in a point, *K*', with coordinates (1+*x*−*y*,1) —where *B* has been taken as the origin of the corresponding Cartesian coordinate system. Hence, the Euclidean distance from *K* to *K*' would be $(1 - \mathrm{y})\sqrt{2}$, from which it follows that the *sum* of all the distances is $\int_0^1 \int_0^x \sqrt{2}\,(1 - \mathrm{y})\,dxdy = \sqrt{2}\int_0^1 \left(\mathrm{x} - \frac{x^2}{2}\right)dx = \sqrt{2}\left(\frac{1}{2} - \frac{1}{6}\right) = \frac{\sqrt{2}}{3}$, which coincides with the product of the area of the triangle *BCF* (the *number* of points in *BCF*), $\frac{1}{2}$, and the length of *RQ*, the segment of lifeline that goes from the barycentre of *BCF* to *FG*, $\frac{2\sqrt{2}}{3}$ (see next footnote).

[12] Indeed, taking B as the origin of a Cartesian coordinate system and using that the Lexis squares have unit sides, we have that the coordinates of the points C, F and G are, respectively, (1,0), (1,1) and (2,1) and that, as a consequence, the coordinates of the barycentres R and S are $\left(\frac{2}{3} - \frac{1}{3}\right)$ and $\left(\frac{5}{3} - \frac{2}{3}\right)$, respectively. Hence, it is not difficult to see that the distance from R to Q is equal to the length of the

here, taking into account that the distances of P to Q and of U to T are equal to $\sqrt{2}$ equivalent to a year, it follows that on average the exposure to risk of each immigrant/emigrant would be, respectively, $\frac{2}{3}$ and $\frac{1}{3}$. Hence bearing in mind the previous arguments, we have that the number of deaths observed in the parallelogram BCFG, $\frac{1}{2}(\mathrm{D_x^t} + \mathrm{D_x^{t+1}})$, could be broken down in the following way:

$$\frac{1}{2}(D_x^t + D_x^{t+1}) \approx \left(C_{x-1}^t - \frac{1}{2}D_{x-1}^t + \frac{1}{2}N_{x-1}^t\right)q_x + \frac{1}{2}N_x^t\frac{2}{3}q_x + \frac{1}{2}N_x^{t+1}\frac{1}{3}q_x$$

And consequently, an estimator for q_x, under HODS, would be obtained by way of the following expression:[13]

$$\tilde{q}_x = \frac{\frac{1}{2}(D_x^t + D_x^{t+1})}{C_{x-1}^t - \frac{1}{2}D_{x-1}^t + \frac{1}{2}N_{x-1}^t + \frac{1}{3}N_x^t + \frac{1}{6}N_x^{t+1}} \qquad (3)$$

3. Comparative analysis

In order to analyse the impact of considering, or not, migration on the estimates of the probability of survival or death at each age x, we have constructed, using the deaths of two adjacent years, the life tables of the years 2006, 2007 and 2008 (for ages 0 to 99 years). The information handled comes from the data that INE offers directly to the public on its website. Sex and age (January, 1) Population Now Cast (ePOBa) estimates (for years 2006, 2007 and 2008) have been used as population data (INE, 2012). Death statistics come from sex and age vital statistics (INE, 2010a). And, approximations to sex and age annual immigrant and emigrant figures (for years 2006 to 2009) have been obtained from the data of residential variations (INE, 2010b)[14]. This section shows the

hypotenuse of a right-angled triangle with right sides both of length $\frac{2}{3}$ and that the segment ST is the hypotenuse of a right-angled triangle of right sides $\frac{1}{3}$.

[13] In this case, the formulae used for ages zero and one have been, respectively, $\tilde{\mathrm{q}}_0 = \frac{0.7\mathrm{D_0^t} + 0.3\mathrm{D_0^{t+1}}}{\mathrm{B^t} + \frac{4}{10}\mathrm{N_0^t} + \frac{1}{10}\mathrm{N_0^{t+1}}}$ and $\tilde{\mathrm{q}}_1 = \frac{\frac{1}{2}(\mathrm{D_1^t} + \mathrm{D_1^{t+1}})}{\mathrm{C_0^t} - 0.3\mathrm{D_0^t} + \frac{1}{2}\mathrm{N_0^t} + \frac{1}{3}\mathrm{N_1^t} + \frac{1}{6}\mathrm{N_1^{t+1}}}$; where for the estimation of $\tilde{\mathrm{q}}_0$ it has been assumed that on average the probability of death of a migrant of age zero located in the lower triangle of the period t is four times the probability of death of a migrant of the same age located in the upper triangle of the period $\mathrm{t} + 1$.

[14] It should be noted that the statistic of residential variations cannot be observed as a completely true source for migration flows given that this is just an account of the entrants and exits registered on the

differences obtained for the 2007 tables, before and after adjusting estimated crude probabilities. The adjustment has been carried out using nonparametric estimation; in particular, through a Gaussian kernel graduation (see, e.g., Ayuso *et al.*, 2007, pp. 217-22). Comparisons between values obtained for each age x with HCDS and HODS were carried out by use of two indicators of dissimilarity widely employed in the literature:

- Absolute relative error (ARE):

$$\frac{|\hat{q}_x - \tilde{q}_x|}{\hat{q}_x}, x = 0,1, \dots, 99.$$

- Square relative error (SRE):

$$\frac{(\hat{q}_x - \tilde{q}_x)^2}{\hat{q}_x}, x = 0,1, \dots, 99.$$

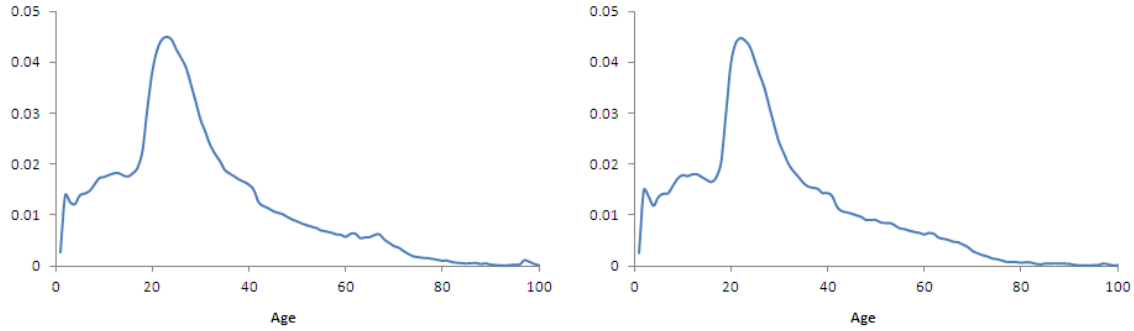

Figure 3. Differences in crude probabilities with and without migration flows: men (left) and women (right).

Figure 3 shows, in graphic form, ARE values obtained by comparing the estimates of crude probabilities achieved for men (left) and women (right), after applying equations (2) and (3), with and without migratory flows. As can be seen, the differences of considering HCDS or HODS are significant, reaching the greatest dissimilarities in the range of 14 to 36 years, with the maximum in both cases being reached at age 22, where the difference is close to 4.5%. Similar results are reached when SRE is used as measure of dissimilarity: the age range with greater differences remains the same.

lists of the municipalities. This has been used, nevertheless, because it represents the only public source that can be used as a proxy of the migration flows occurring in Spain during the period analyzed.

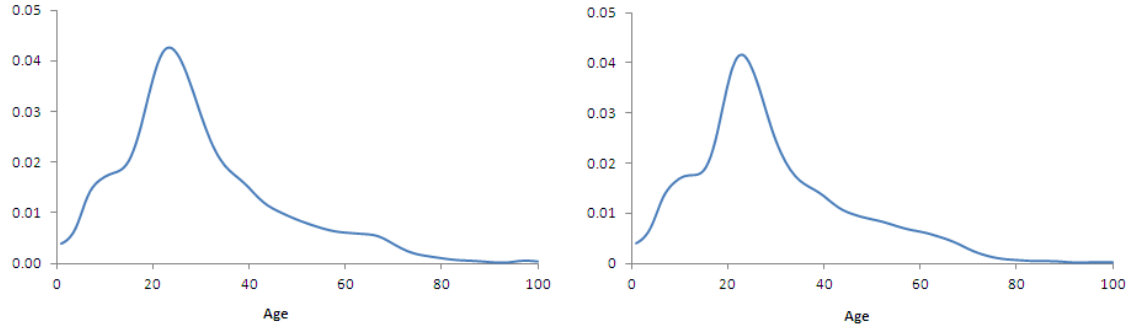

Figure 4: Differences in graduated probabilities with and without migratory flows: (left) men and (right) women.

The discrepancies observed reveal, at least in this case, the usual assumption of HCDS being inadequate. This provokes a systematic, not uniform overestimation in the probabilities of death for all ages; which, as can be seen below (Table 1 and Figure 5), has asymmetrical effects on the results of different actuarial calculations.

In actuarial calculations, however, the crude death estimates obtained directly from observed data, $\hat{q}_x$ or $\tilde{q}_x$, are not usually used without being graduated first. The objective of graduation is to soften the crude estimates in a way that eliminates (or mitigates) the random fluctuations present in empirical data. In this study, graduation has been carried out using a kernel graduation (see, e.g., Ayuso *et al.*, 2007). In particular, the kernel estimation carried out uses a Gaussian function as a kernel with a window parameter, or bandwidth, equal to 1.[15]

Once the initial values $\hat{q}_x$ and $\tilde{q}_x$, were graduated, the indicators of dissimilarity ARE and SRE introduced previously were again calculated. The results obtained for men and women with the ARE measurement are shown in Figure 4. The comparison with the graduated probabilities does not change in any way the conclusions reached previously; in fact, they serve to reinforce the results already obtained.

Finally, in line with Pavía and Escuder (2003), some specific probabilities have been obtained with the aim of illustrating the differences that could be derived by using one or other hypothesis on the demographic system: Table 1 shows the results. As was expected of a demographic situation such as that lived in Spain, where in recent years entries have been significantly greater than exits, the non-inclusion of migratory flows underestimates the survival probabilities and overestimates the death probabilities.

[15] Similar results were obtained with alternative bandwidth parameters.

Differences in every case are evident to the third significant digit in men and (at most) the fourth digit in women. The impact, therefore, is different depending on the gender. In spite of the repercussions on the individual probabilities being similar for both sexes (see Figures 3 and 4), the inclusion of migratory flows has a significantly greater impact on men than on women, at least for the range of ages and periods considered.

Table 1: Examples of probabilities.

	Men		Women	
	HCDS	HODS	HCDS	HODS
${}_{25}q_{40}$	.1380204	.1371103	.0584317	.0580243
${}_{15\|10}q_{50}$	.2345362	.2338640	.1112529	.1109601
${}_{85}P_{0}$	.3442274	.3452564	.5672970	.5679357
${}_{20}P_{15}$	.9872419	.9876434	.9951881	.9953219

This asymmetric impact is also clearly visible in Figure 5, where the differences in life expectancy when either migration flows have been or have not been taken into account are shown. As can be observed, the underestimation in life expectancy that entails the non-inclusion of migratory flows is, almost for all ages, double in men than in women.

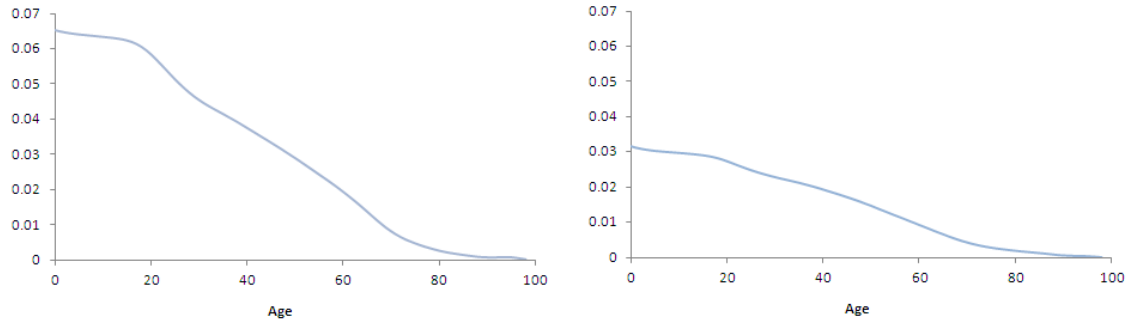

Figure 5: Differences in life expectancy (in years) with and without migration flows: (left) men and (right) women.

4. Conclusions

When working with general populations, the usual practice in the construction of life tables consists in ignoring the entry and exit flows that occur in the study group during the analysis period, under the assumption that these usually have little value compared to the size of the population. The use of a closed demographic system hypothesis has been consequently the norm among analysts.

In this paper, (i) the techniques used for estimating death probabilities have been extended to open demographic systems with aggregate data; and, (ii) the resulting estimator has been used, along with the classic HCDS estimator, to obtain (from 0 to 99 years) life tables of the resident population in Spain from 2006 to 2008.

Comparison of crude and graduated probabilities obtained with and without the inclusion of migratory flows shows the impact that entry and exit movements can have on actuarial and demographic calculations. In the examples considered the repercussion has been asymmetric by age and sex. On one hand, the greatest discrepancies, in relative terms, are concentrated in the range of ages from 14 to 36 years, where the intensity of flows has been stronger in Spain in the recent years. On the other hand, by gender, it is clear that the impact on women is less, in contrast to that in men. The well-known lower probabilities of death that women suffer in the range of ages where greater probabilities of migration occur may be behind this result.

The results obtained in this paper point to the need to explicitly consider migratory flows in the estimation of life tables for general populations. The cost to introduce this information is minimal but the potential danger could be significant, especially in situations where entry movements are well above exit movements. In this type of situation, to omit migration flows would lead to an overestimation of probabilities of death and hence to an underestimation of life expectancy, with the danger that this could entail for a correct inter-generational planning that would ensure an adequate stability of the social security pension programmes characteristic of the Western welfare systems.

Certainly, beyond the possible influence that migratory flows and other relevant information might have on results, the great quantity of data provided by modern statistical systems offers an opportunity to seek new ways to exploit the available data in more efficient fashions. So, developing new methodological approaches or implementing proper analyses that help to assess the soundness of broadly used hypotheses should be included early on the statistical demographic research agenda. Along this line, in order to assess the cumulative impact of migration in mortality in Spain, it would be interesting to compare the probabilities of death and life expectancies of born-in-Spain and total (Spaniards and foreigners) populations. Likewise, death and

migrant microdata should be analysed in order to ascertain the suitability of the uniform distribution hypotheses required when handling aggregate data.

Assessing Implicit Hypotheses in Life Table Construction (A2)

Josep Lledó, Universitat de Valencia

Jose M. Pavía, Universitat de Valencia

Francisco G. Morillas, Universitat de Valencia

Abstract

Mortality figures are of capital importance for policy-making and public planning, as in forecasting financial provisions in public pension systems. General population life tables are constructed from aggregated statistics, an issue that usually entails adopting some (implicit) assumptions in their construction, such as the hypothesis of closed demographic system or the hypotheses of uniform distributions of death counts (and migration events) by age and calendar year. As microdata have become more abundant and reliable, these hypotheses could be assessed and more assumption-free estimators employed. Using a real database from Spain, we show that the above hypotheses are not appropriate in general and that the more efficient estimators proposed in this paper should be promoted, as differences persist depending on the estimator computed.

Key Words: mortality tables; death probabilities; migration flows; cohort-based estimators; microdata; point patterns; efficiency.

Acknowledgements

The authors wish to thank two anonymous referees for their valuable comments and suggestions and Instituto Nacional de Estadistica for their first-rate assistance in responding to our request for the detailed statistics of deaths handled in this research. Thanks are also due to M. Hodkinson for reviewing the English of the paper. The usual disclaimer applies. This work was supported by the Spanish Ministry of Economics and Competitiveness under grant CSO2013-43054-R.

1 Introduction

Mortality figures are capital for public planning and policy-making. They are routinely employed in many countries to generate demographic projection scenarios (Eurostat, 2010) and financial provision forecasts. In the US they are used to obtain yearly estimates of the future costs of the Old-Age, Survivors, and Disability Insurance federal programme (Soneji and King 2012). Mortality is commonly represented in the form of a life table, which shows the probabilities that the members of a particular population have of living to, $\mathrm{p_x}$, or dying within one year, $\mathrm{q_x} = 1 - \mathrm{p_x}$, at each exact age x. The life expectancy (at birth), $\mathrm{e_0}$, is probably the most popular statistic of a life table.

Statisticians from official statistical agencies are in charge of estimating national life tables. In a classical, Laplacian interpretation of probability, the proportion between the number of deaths and the number of people at risk of dying (within a given population group and time period) provides initial (crude) estimates for the probability of death in each age group, $\mathrm{q_x}$. These probabilities can also be estimated from the rate between the number of deaths and the average population at risk of dying[16], $\mathrm{m_x}$. These probabilities are subjected to graduation or adjustment processes to smoothing the profile of the associated stochastic process.

General population life tables are usually computed from aggregated statistics – 'birth and death counts from vital statistics, plus population counts from periodic censuses and/or official population estimates' (Wilmoth *et al.,* 2007, p. 1), which entails adopting assumptions to construct them. Supposing a uniform distribution of the annual tabulations of death counts by age, not explicitly considering migratory flows, or even directly accepting the hypothesis of closed demographic systems are among the most common hypotheses. The specific (implicit) hypotheses adopted depend on the particular estimator used. For example, the hypothesis of uniform distribution of birthdates (by age and sex) is required in period-based (m-type) estimators (INE, 2009; ONS 2010, 2012), whereas it is not necessary in (biannual-period) cohort-based (q-type) estimators (INE 2007; Pavia *et al.,* 2012).

[16] Depending on the estimator used and/or the literature, the denominator of m_x is also computed/defined as the central exposed to risk population or the total number of 'person-years' at risk.

As actual microdata have become more abundant and more reliable, some of the most frequently employed hypotheses could be assessed and more assumption-free estimators constructed. This research moves along this line. In particular, we delve into the issue of introducing detailed data in the problem of estimating probabilities of death from official statistics. In this sense, we follow Muriel de la Riva *et al.* (2010) - who point out that "a major effort of research and methodological innovation should open the way to optimal utilisation of the available data on deaths" (p. 1) - and Pavía *et al.* (2012) - who, after showing the influence of omitting migratory flows, challenge scholars to develop new methodological approaches which exploit the great quantity of data currently provided by statistical systems and encourage analysts to carry out proper studies to assess the (implicit) hypotheses used in life table construction.

In this paper, we estimate and compare several life tables under different levels of information available and we assess in the process the hypotheses assumed in the less detailed levels. More specifically, starting with a situation where only aggregated annual death and population counts (by age and gender) are known, we gradually enhance the levels of detailed information available until a state is reached in which both death and migrant microdata are accessible. The study is performed using a real database going deeper in the family of cohort-based estimators proposed in Pavia *et al.* (2012).

Compared to the currently more popular period-based estimators[17], the (biannual) cohort-based estimators have the advantages of not needing (i) to assume the assumption of uniform distribution of birthdates (an untestable hypothesis with our data) and (ii) to pool deceased figures from different cohorts (generations); although in contrast they combine deaths occurring in different calendar years. Additionally, they also benefit from not violating the principle of correspondence (all favorable cases are possible cases) and from reflecting the experience of a real group of people (Hinde, 1998, p. 16).

There is a vast literature on measuring and modelling mortality risks. Fuelled by management applications in longevity risks and the provision of pensions (e.g. Cairns *et*

[17] For example, estimators of this family are currently used, although with great differences, in Spain (INE, 2009) and in the UK (Hinde, 1998; ONS, 2010, 2012).

al., 2008; Cairns, 2013; Rodriguez-Pardo del Castillo *et al.,* 2015), the field of stochastic mortality modelling has seen a rapid growth in recent years. Indeed, building on the work of Lee and Carter (1992), we are experiencing a real explosion in research that seeks to offer a more accurate description of the underlying patterns of mortality improvements (e.g. Lazar and Denuit, 2009; Russolillo *et al.,* 2011; Danesi *et al*., 2015; Currie, 2016). Nevertheless, the accuracy of model forecasts depends, among other issues, on the quality of the input data and on the conceptual framework assumed. We focus on introducing detailed data in the process of estimating input (raw) data from official statistics. Other approaches include considering life style, environment and advances in medicine as risk factors (Woo *et al.,* 2009), the analysis of cohort effects by cause of death (Willets, 2004) or the increase in efficiency and accuracy by pooling and tying data from different groups (Li and Lee, 2005). Ahrems and Pigeot (2007) offer some examples from an epidemiological perspective.

The rest of the paper is structured as follows. Section 2 is devoted to methodological issues: we present the terminology used and the particular formulae and hypotheses employed to convert raw data into crude mortality probability estimates. In section 3, we assess the usual implicit hypothesis of closed demographic system and of uniform distributions of deaths and migrants as well as their materializations in the mathematical expressions proposed. In section 4, we apply to a real dataset the formulae introduced and compare the different life tables obtained after scrutinizing further into the level of detailed information employed. Finally section 5 concludes the paper. A graphical appendix (provided as supplementary material) complements the paper.

2. Methodology

The particular formulae and hypotheses used to convert raw data into crude mortality probability estimates are introduced in this section. To make it easier to understand the notation (Table 1) and equations employed (Table 2), we rely on the Lexis diagram and the use of lifelines (see, e.g. Caselli *et al.*, 2006; Wilmoth *et al.,* 2007) to represent geometrically the available data of our target population.

Table 1. Detail of symbols used in the equations.

Notation	Description
x	Age, measured in years.
t	Calendar year, measured in years.
q_x	Probability of death during the age interval x to $x+1$.
m_x	(Central) rate of mortality at age x.
C_x^t	Population with completed age (age at last birthday) x in January 1 of year t.
D_x^t	Number of deaths in year t with completed age x.
E_x^t	Number of emigrants in year t with completed age x.
I_x^t	Number of immigrants in year t with completed age x.
N_x^t	Net migrants in year t with completed age x.
B^t	Births in year t.
$D_{x:r}^t$	Number of deaths in year t with completed age x born in year r.
$E_{x:r}^t$	Number of emigrants in year t with completed age x born in year r.
$I_{x:r}^t$	Number of immigrants in year t with completed age x born in year r.
$D_{x:t-x}^{t,t+1}$	Number of deaths with completed age x (during years t and $t+1$) of cohort $t-x$.
$E_{x:t-x}^{t,t+1}$	Number of emigrants with completed age x (during years t and $t+1$) of cohort $t-x$.
$b_{x,j}^t$	Years lived with completed age x for the jth deceased person in year t.
$e_{x,j}^t$	Years lived with completed age x in the target population for the jth emigrant of year t.
$i_{x,j}^t$	For the jth immigrant of year t with completed age x, the difference in years between its $x+1$ birth date and its immigration date.

The Lexis diagram is a device for depicting the stock and flow of a population and the occurrence of demographic events over calendar time and age. It is a (time, age) coordinate system in which individual lives are represented by line segments (lifelines) of unit slope, which join (time, age of) birth or immigration and (time, age of) death or emigration events. In each particular year t and age *x*, an individual lifeline may end in death, denoted by '×' (lines a and b in Figure 1-left), or emigration, denoted by '□' (line c). An individual may also immigrate into the population, denoted by '○' (lines d and e), or merely pass through the section of the Lexis diagram under study (lines f and g).

Depending on the level of data available, the lifelines of each individual and/or the points in the diagram where demographic events occur are known exactly or only approximately. When only counts of the number of individuals who are alive (by age) and counts (by year and age) of deaths (and migrants) are accessible, we adopt the usual convention of the Lexis scheme of assigning values to segments and surfaces. The value of a segment is always identified by the number of lifelines that crosses it, while the value(s) to be associated with each area will be identified with the number of death (or immigrant or emigrant) events occurring within it. For example, the value of the segment AB of Figure 1-left is identified with the number of residents counted (estimated) at age $\mathrm{x}-1$ in January 1 of year t, $\mathrm{C}_{\mathrm{x}-1}^{\mathrm{t}}$. More specifically, we denote by $\mathrm{C}_{\mathrm{x}}^{\mathrm{t}}$ the stock of (counted or estimated) population with completed age x on 1 January of

year t (segment BE in Figure 1), by D_x^t the number of deaths with completed age x registered in year t (i.e. the number of crosses, '×' symbols, within BCEF, Figure 1-right), by E_x^t the number of individuals with completed age x leaving the target population (emigrants) in year t (i.e. the number of squares, '□' symbols, within BCEF, Figure 1-right), by I_x^t the number of individuals with completed age x coming to the population (immigrants) in year t (i.e., the number of circles, 'o' symbols, within BCEF, Figure 1-right), by $N_x^t = I_x^t - E_x^t$ the net migration with completed age x recorded in year t and by B^t the number of births in year t (the baseline of the Lexis diagram corresponding to year t).

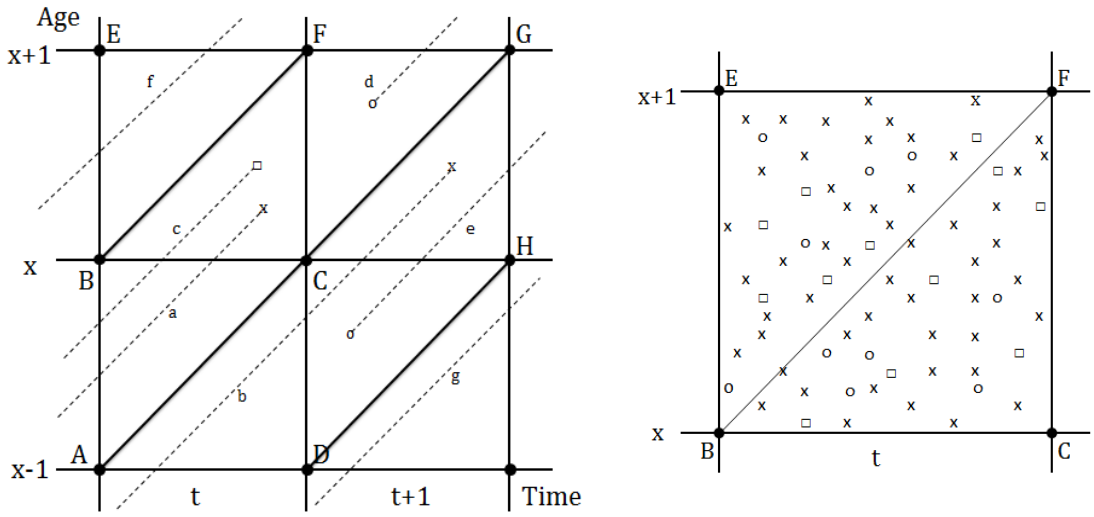

Figure 1. Small section of Lexis diagram with some lifelines (left) and schematic representation of death (×) and migrant (□, ○) events in a 1x1 cell (right).

Sometimes, the information is available at a slightly more detailed level and, in addition to by age and year of event (i.e. period), death and (maybe) migrant counts are also available by year of birth (i.e. cohort). In such cases, the counts (which are represented by a Lexis triangle) are denoted in a different way to reflect the extra-information available. In particular, we denote by $D_{x:r}^t$ the number of deaths with completed age x in year t born in year r (i.e. the number of crosses inside triangle BCF of Figure 1-right when $r = t - x$ and the number of crosses inside triangle BEF if $r = t - x - 1$), by $E_{x:r}^t$ the number of emigrants with completed age x in year t born in year r (i.e. the number of squares inside triangles BCF and BEF of Figure 1-right when $r = t - x$ and $r = t - x - 1$, respectively) and by $I_{x:r}^t$ the number immigrants with completed age x in year t born in year r (i.e. the number of circles inside triangle BCF of Figure 1-right if

$r = t - x$ and the number of circles inside triangle BEF when $r = t - x - 1$). Likewise, we denote by $D^{t,t+1}_{x:t-x} = D^{t}_{x:t-x} + D^{t+1}_{x:t-x}$, $E^{t,t+1}_{x:t-x} = E^{t}_{x:t-x} + E^{t+1}_{x:t-x}$ and $I^{t,t+1}_{x:t-x} = I^{t}_{x:t-x} + I^{t+1}_{x:t-x}$, respectively, the number of deaths, emigrants and immigrants with completed age x of cohort $t - x$ (i.e. the number of crosses, squares and circles, respectively, located inside the parallelogram BCFG).

When individual dates of birth and of the event of death or migration are available, crosses, squares and circles can be accurately located in the Lexis diagram and a better approach of the populations at risk attained. In this scenario, we denote by $b^{t}_{x,j}$ the difference between the date of death and the date of last birthday of the jth deceased person with completed age x in year *t*. For example, looking at crosses L and J in Figure 2-left (and identifying point-events and individuals with the same symbol to alleviate notation), we have that $b^{t}_{x,L}$ and $b^{t}_{x,J}$ are, respectively, equal to the lengths of the segments LN and JK. Observe that when the cross is located in the upper triangle (point L in Figure 2-left) $b^{t}_{x,L}$ could be also computed as one minus the distance between L and N' and equals one minus its date of death and the date of its birthday in year *t*. Similarly, we denote by $e^{t}_{x,j}$ the difference between the dates of emigration and last birthday of the jth emigrated person in year t with completed age *x*. As happens with $b^{t}_{x,j}$, this quantity coincides with the time lived (in years) with completed age x for the individual as a member of the target population. Examples of emigrants are individuals X and O in Figure 2-right. Finally, we denote, for the jth immigrant with completed age x in year t, by $i^{t}_{x,j}$ the difference between its $x + 1$ birth date and its immigration date. Note that $i^{t}_{x,j}$ has been defined in a reversed way to $b^{t}_{x,j}$ and $e^{t}_{x,j}$. For example, for the immigrants with immigration events denoted by R and U in Figure 3-right, $i^{t}_{x,R}$ and $i^{t}_{x,U}$ are, respectively, the distance of the event point to segment EF. That is, the lengths of the segments RT and UW.

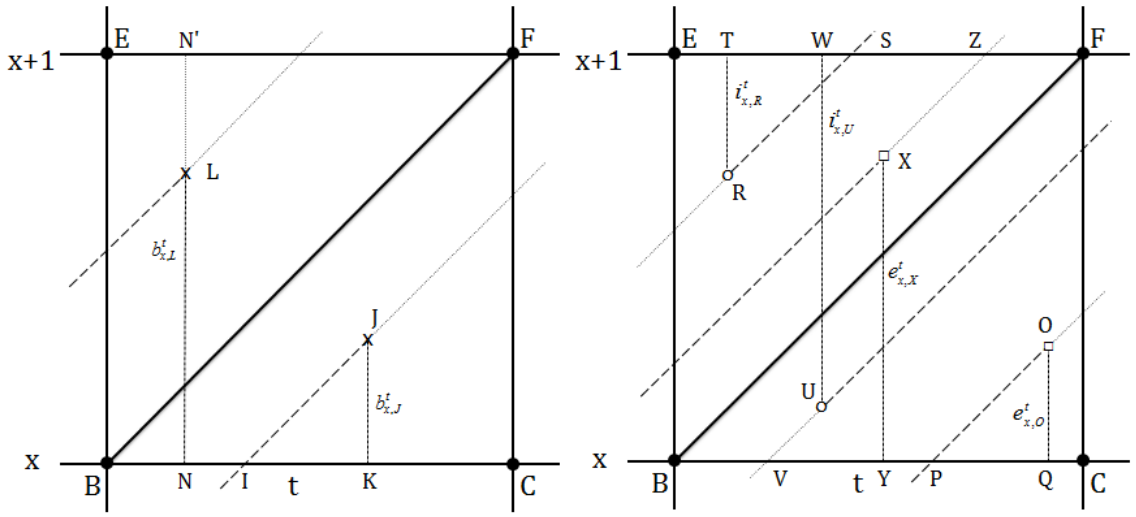

Figure 2. Detail in the Lexis scheme for some death (left) and migrant events (right). Notes. $b^t_{x,L}$, $b^t_{x,J}$ $e^t_{x,X}$ and $e^t_{x,O}$ measure the distance between either the events of death (L and J) or emigration (X and O) of the corresponding individuals and the dates of their x birthdays. $i^t_{x,R}$ and $i^t_{x,U}$ measure, respectively, the distances between the events of immigration of U and R and their corresponding dates of x + 1 birthdays. Except in exceptional cases (such as when a person immigrates and dies, immigrates and emigrates or emigrates and immigrates with the same age), these quantities account for the exact time (in years) exposed to the risk of dying with completed age x as a member(s) of the target population for the individuals identified by points J, L, R, X, U and O. Point-events and individuals are identified with the same symbol to alleviate notation.

The above definitions will allow one to easily calculate, under each level of available information and corresponding (implicit) hypotheses, the contribution of each individual to both initial and central exposed-to-risk populations and from them obtaining, by age (and sex), crude probability and rate estimates of mortality, q_x and m_x (see Table 2). Disposing of both (probability and rate) estimates will allow comparison in some scenarios of their theoretical relationship.

Table 2. Summary of estimators and hypotheses used to derive them.

Estimator	Hypotheses		
	Closed demographic system	Uniform	
		Deaths	Migrants
$\hat{q}_x = \frac{\frac{1}{2}(D_x^t + D_x^{t+1})}{C_{x-1}^t - \frac{1}{2}D_{x-1}^t}$ $\hat{m}_x = \frac{\frac{1}{2}(D_x^t + D_x^{t+1})}{C_{x-1}^t - \frac{1}{2}D_{x-1}^t - \frac{1}{3}D_x^t - \frac{1}{6}D_x^{t+1}}$	Yes	Yes	N/A
$\tilde{q}_x = \frac{\frac{1}{2}(D_x^t + D_x^{t+1})}{C_{x-1}^t - \frac{1}{2}D_{x-1}^t + \frac{1}{2}N_{x-1}^t + \frac{1}{3}N_x^t + \frac{1}{6}N_x^{t+1}}$ $\tilde{m}_x = \frac{\frac{1}{2}(D_x^t + D_x^{t+1})}{C_{x-1}^t - \frac{1}{2}D_{x-1}^t + \frac{1}{2}N_{x-1}^t + \frac{1}{3}N_x^t + \frac{1}{6}N_x^{t+1} - \frac{1}{3}D_x^t - \frac{1}{6}D_x^{t+1}}$	No	Yes	Yes
$\breve{q}_x = \frac{D_{x:t-x}^t + D_{x:t-x}^{t+1}}{C_{x-1}^t - D_{x-1:t-x}^t} = \frac{D_{x:t-x}^{t,t+1}}{C_{x-1}^t - D_{x-1:t-x}^t}$ $\breve{m}_x = \frac{D_{x:t-x}^t + D_{x:t-x}^{t+1}}{C_{x-1}^t - D_{x-1:t-x}^t - D_{x:t-x}^{t,t+1} + \sum_{j=1}^{D_{x:t-x}^t} b_{x,j}^t + \sum_{j=1}^{D_{x:t-x}^{t+1}} b_{x,j}^{t+1}}$	Yes	No	N/A
$\ddot{q}_x = \frac{D_{x:t-x}^t + D_{x:t-x}^{t+1}}{\ell_x^t - E_{x:t-x}^{t,t+1} + \sum_{j=1}^{E_{x:t-x}^t} e_{x,j}^t + \sum_{j=1}^{E_{x:t-x}^{t+1}} e_{x,j}^{t+1} + \sum_{j=1}^{I_{x:t-x}^t} i_{x,j}^t + \sum_{j=1}^{I_{x:t-x}^{t+1}} i_{x,j}^{t+1}}$ $\ddot{m}_x = \frac{D_{x:t-x}^t + D_{x:t-x}^{t+1}}{R_x^t - D_{x:t-x}^{t,t+1} + \sum_{j=1}^{D_{x:t-x}^t} b_{x,i}^t + \sum_{j=1}^{D_{x:t-x}^{t+1}} b_{x,i}^{t+1}}$	No	No	No

$\ell_x^t = C_{x-1}^t - D_{x-1:t-x}^t + N_{x-1:t-x}^t$ and $R_x^t = \ell_x^t - E_x^{t,t+1} + \sum_{j=1}^{E_{x:t-x}^t} e_{x,j}^t + \sum_{j=1}^{E_{x:t-x}^{t+1}} e_{x,j}^{t+1} + \sum_{j=1}^{I_{x:t-x}^t} i_{x,j}^t + \sum_{j=1}^{I_{x:t-x}^{t+1}} i_{x,j}^{t+1}$. N/A: Not applicable.

2.1. Closed demographic system and uniform distribution of deaths by age and calendar year

The most common and less data-demanding scenario consists of assuming a closed demographic system and a uniform distribution of deaths by age and calendar year. For this classical situation, estimators have been proposed within the cohort-based family. As has been stated in the introduction the family of estimators proposed in Pavia *et al.* (2012) is followed in this work. Compared to other estimators also sharing the same approach, such as the one suggested in INE (2007), these estimators have the advantage of not masking the impact of the omission of migratory flows, making it easier to assess their actual effect on mortality estimates. In particular, for $x = 2,3, \ldots, \omega$ (where ω is the maxim age after which nobody can survive), the estimators employed for q_x and m_x are:

$$\hat{q}_x = \frac{\frac{1}{2}(D_x^t + D_x^{t+1})}{C_{x-1}^t - \frac{1}{2}D_{x-1}^t} \quad (1)$$

$$\hat{m}_x = \frac{\frac{1}{2}(D_x^t + D_x^{t+1})}{C_{x-1}^t - \frac{1}{2}D_{x-1}^t - \frac{1}{3}D_x^t - \frac{1}{6}D_x^{t+1}} \quad (2)$$

which, under the hypothesis of uniform distribution of deaths, must verify:

$$\widehat{m}_x = \frac{\hat{q}_x}{1 - \frac{1}{2}\hat{q}_x} \tag{3}$$

Using the assumed hypotheses the above expressions can be easily derived from the geometry of the Lexis diagram. Under the usual convention of the Lexis scheme of assigning values to segments and surfaces, the segment BC will represent the number of people reaching age x in year t and the quadrilateral BFGC will represent the number of them who die before reaching age *x+1*. Hence, given that q_x is the probability that someone aged exactly x die before reaching age *x+1*, we have that the estimator for q_x from which (1) comes $\frac{\mathrm{BFGC}}{\mathrm{BC}}$. On the one hand, $\mathrm{BFGC} = \mathrm{BCF} + \mathrm{CFG}$, which summands under the hypothesis of uniform distribution of deaths by age and calendar year are equal to half of $\mathrm{BEFC} = D_x^t$ and $\mathrm{CFGH} = D_x^{t+1}$, respectively. On the other hand, $\mathrm{BC} = \mathrm{AB} - \mathrm{ABC} = \mathrm{AB} - \frac{1}{2}\mathrm{ABCD} = C_{x-1}^t - \frac{1}{2}D_{x-1}^t$.

Deriving (2) requires a few more calculations. To compute the total time exposed to the risk of dying between ages x and $x+1$ for the people counted in BC, we divide out these into three groups: (i) those that reach age $x+1$, each of which live a whole year, and are $\mathrm{FG} = \mathrm{BC} - \mathrm{BFGC}$; (ii) those who die in year t, each of which live in average $\frac{1}{3}$ of year before dying, and are $\mathrm{BCF} = \frac{1}{2}\mathrm{BEFC}$; and (iii) those who die in year $t+1$, each of which live in average $\frac{2}{3}$ of year before dying, and are $\mathrm{CFG} = \frac{1}{2}\mathrm{CFGH}$.[18] Hence, the denominator of (2) equals:

$$\mathrm{BC} - \mathrm{BFGC} + \frac{1}{3}\cdot\frac{1}{2}\mathrm{BEFC} + \frac{2}{3}\cdot\frac{1}{2}\mathrm{CFGH} = C_{x-1}^t - \frac{1}{2}D_{x-1}^t - \frac{1}{2}D_x^t - \frac{1}{2}D_{x+1}^{t+1} + \frac{1}{6}D_x^t + \frac{2}{6}D_{x+1}^{t+1} =$$
$$C_{x-1}^t - \frac{1}{2}D_{x-1}^t - \frac{1}{3}D_x^t - \frac{1}{6}D_x^{t+1}.$$

[18] Under the hypothesis of uniform distribution of deaths by age and calendar year, the average lifespan of those dying within triangles BCF and CFG can be calculated as follows. Taking B as the origin of a Cartesian coordinate system and considering an arbitrary point P with Cartesian coordinates (t_1, t_2) within BCF, it is not difficult to prove that the lifeline of P intersects BC in a point P' with coordinates $(t_1 - t_2, 0)$, from which follows that the Euclidean distance from *P* to *P'* is $\sqrt{2}t_2$ and that the *sum* of all the distances of those dying within BCF is $\int_0^1 \int_0^{t_1} \sqrt{2}t_2 dt_2 dt_1 = \frac{\sqrt{2}}{6}$. Dividing this sum by the Euclidean area of BCF gives us an average distance of $\frac{\sqrt{2}}{3}$ in lifeline distance, which joined to the fact that a Euclidean distance of $\sqrt{2}$ of a lifeline corresponds to a year yields $\frac{1}{3}$ of year as average of the time living with age x for those dying within BCF. A similar reasoning leads to the number $\frac{2}{3}$ for those dying within CFG. An alternative proof can be found in Appendix A of Carstensen (2007).

Despite the uniform distribution of deaths by age and calendar year assumed, nobody would maintain the uniform assumption for zero years. As it is well-known, deaths of children less than one year old are concentrated in their first weeks of life. Therefore, different expressions are routinely used for ages zero and one[19]. Thus, following Pavía *et al.* (2012), the formulae for ages zero and one in this scenario would be:[20] $\hat{q}_0 = \frac{0.7D_0^t + 0.3D_0^{t+1}}{B^t}$, $\hat{q}_1 = \frac{\frac{1}{2}\left(D_1^t + D_1^{t+1}\right)}{C_0^t - 0.3D_0^t}$, $\widehat{m}_0 = \frac{0.7D_0^t + 0.3D_0^{t+1}}{\mathrm{B}^t - \frac{1}{2}D_0^t - \frac{1}{6}D_0^{t+1}}$ and $\widehat{m}_1 = \frac{\frac{1}{2}\left(D_1^t + D_1^{t+1}\right)}{C_0^t - 0.3D_0^t - \frac{1}{3}D_1^t - \frac{1}{6}D_1^{t+1}}$.

2.2. Open demographic system and uniform distribution of deaths and migrants

As it is well known, migration plays a key role in population change and growth. Despite this, in this era of high migration (Ediev *et al.*, 2014) the explicit consideration of migration flows has been systematically ignored in life table construction. However, as shown in Pavia *et al.* (2012), omitting migratory flows has an asymmetrical repercussion by age and sex and leads to a misestimating of probabilities of death; an issue that could be potentially dangerous for pension system planning when entry movements are well above exit flows.

In this scenario, adjustment does need to be made to the exposed-to-risk populations to take account of those persons who, because of international movements, are not members of the target population during the period window. First, to compute BC (from AB) we need to subtract those emigrating in ABC and to add those immigrating

[19] Special treatments are also advisable for advanced ages, a matter of great interest because progressively more and more people are living to very high ages. In these cases, data are scarcer, and also less reliable (Cairns *et al.*, 2016, forthcoming). Hence, researchers are forced to use models and to pool data from several calendar periods, countries and/or ages to reach reliable estimates (Society of Actuaries, 2005). Although this topic deserves to be dealt with in more detail, we will not pursue this issue here for not obscuring, with the associated complexities, the focus of this research.

[20] Although nowadays (almost) all statistical systems record at least the cohort for aged zero deaths, in order to be completely respectful with the hypotheses of this scenario we will consider that only statistics of the year of decease are recorded. Under these circumstances, in Spain they have been traditionally distributed 70% in the lower triangle and 30% in the upper one (Goerlich, 2008). According to Pavía *et al.* (2012, p. 107), these numbers could be reached "assuming that the number of deaths occurring during the first half of age zero is approximately four times the number of deaths registered during the second half". This assumption together with the hypothesis of uniform distribution of deaths inside each half of the age also allows obtaining that deaths located in the lower triangle live in average $\frac{1}{4}$ of year and deaths in the upper triangle $\frac{1}{2}$ of year, from which the formulae follow.

in ABC; i.e. the net migration flow in ABC, which by virtue of the uniform hypothesis is $\frac{1}{2}N_{x-1}^{t}$. Likewise, after similar calculations to those performed in footnote 18, we see that those who emigrate within triangles BCF and CFG live an average of $\frac{1}{3}$ years and $\frac{2}{3}$ years as members of the target population and that, conversely, those who immigrate are exposed to risk as members of the target population $\frac{2}{3}$ and $\frac{1}{3}$ years. Hence, as emigrants are initially included in BC, we see that the adjustment to be made to BC due to international movements is $\frac{2}{3}\frac{1}{2}I_x^t + \frac{1}{3}\frac{1}{2}I_x^{t+1} - \frac{1}{2}E_x^t - \frac{1}{2}E_x^{t+1} + \frac{1}{3}\frac{1}{2}E_x^t + \frac{2}{3}\frac{1}{2}E_x^{t+1} = \frac{1}{3}(I_x^t - E_x^t) + \frac{1}{6}(I_x^{t+1} - E_x^{t+1}) = \frac{1}{3}N_x^t + \frac{1}{6}N_x^{t+1}$. In particular, in this scenario the death probability and rate estimators are given, for $x = 2,3 \ldots, \omega$, by equations (4) and (5).

$$\tilde{q}_x = \frac{\frac{1}{2}(D_x^t + D_x^{t+1})}{C_{x-1}^t - \frac{1}{2}D_{x-1}^t + \frac{1}{2}N_{x-1}^t + \frac{1}{3}N_x^t + \frac{1}{6}N_x^{t+1}} \tag{4}$$

$$\tilde{m}_x = \frac{\frac{1}{2}(D_x^t + D_x^{t+1})}{C_{x-1}^t - \frac{1}{2}D_{x-1}^t + \frac{1}{2}N_{x-1}^t + \frac{1}{3}N_x^t + \frac{1}{6}N_x^{t+1} - \frac{1}{3}D_x^t - \frac{1}{6}D_x^{t+1}} \tag{5}$$

which again, due to the hypothesis of uniform distribution of deaths, are related by $\tilde{m}_x \approx \frac{\tilde{q}_x}{1-\frac{1}{2}\tilde{q}_x}$.

Likewise, as in the previous scenario, special expressions have been also employed for ages zero and one: $\tilde{q}_0 = \frac{0.7D_0^t + 0.3D_0^{t+1}}{B^t + \frac{1}{3}N_0^t + \frac{1}{6}N_0^{t+1}}$, $\tilde{q}_1 = \frac{\frac{1}{2}(D_1^t + D_1^{t+1})}{C_0^t - 0.3D_0^t + \frac{1}{2}N_0^t + \frac{1}{3}N_1^t + \frac{1}{6}N_1^{t+1}}$, $\tilde{m}_0 = \frac{0.7D_0^t + 0.3D_0^{t+1}}{B^t - \frac{1}{2}D_0^t - \frac{1}{6}D_0^{t+1} + \frac{1}{3}N_0^t + \frac{1}{6}N_0^{t+1}}$ and $\tilde{m}_1 = \frac{\frac{1}{2}(D_1^t + D_1^{t+1})}{C_0^t - 0.3D_0^t + \frac{1}{2}N_0^t - \frac{1}{3}D_1^t - \frac{1}{6}D_1^{t+1} + \frac{1}{3}N_1^t + \frac{1}{6}N_1^{t+1}}$.

2.3. Closed demographic system with no hypothesis about distribution of deaths

Given that there seems to be certain evidence supporting the idea that, in general, the proportion of deaths in lower and upper Lexis triangles of the same year varies with age (e.g. Vallin, 1973; Preston *et al.*, 2001; Wilmoth *et al.*, 2007)[21], the assumption of uniform distribution of deaths by age and calendar year should not be maintained.

The more broadly followed procedure to challenge this hypothesis has traditionally consisted in collecting deaths by period and cohort - or, in the absence of

[21] In addition to the well-known concentration of deaths in the lower triangle during the first year of life, Wilmoth *et al.* (2007, p. 11) point out that, "*at any age, the distribution of deaths across the two triangles is affected by the relative size of the two cohorts (and sometimes by historical events as well)*", as, for instance, happens with cohorts born at beginnings or ends of major wars (Vallin, 1973, pp. 39-40).

this data, in searching an adequate mechanism to properly split D_x^t (see, e.g. Appendix A in Wilmoth *et al.*, 2007) - to then assume a uniform distribution of deaths inside each triangle. Nowadays, however, thanks to the IT revolution, collecting, storing, and transmitting detailed data have become simpler than ever and likewise handling microdata has been converted into an easy task. In this state of affairs, no hypotheses about the distribution of deaths need to be assumed and new estimators can be developed (see, e.g. INE, 2009). In particular, within the family of estimators handled in this paper and under the hypotheses of this subsection, it follows straightforwardly that, for $x = 1,2,3 \ldots, \omega$, the estimators of q_x and m_x are given, respectively, by:

$$\breve{q}_x = \frac{D_{x:t-x}^t + D_{x:t-x}^{t+1}}{C_{x-1}^t - D_{x-1:t-x}^t} = \frac{D_{x:t-x}^{t,t+1}}{C_{x-1}^t - D_{x-1:t-x}^t} \tag{6}$$

$$\breve{m}_x = \frac{D_{x:t-x}^t + D_{x:t-x}^{t+1}}{C_{x-1}^t - D_{x-1:t-x}^t - D_{x:t-x}^{t,t+1} + \sum_{j=1}^{D_{x:t-x}^t} b_{x,j}^t + \sum_{j=1}^{D_{x:t-x}^{t+1}} b_{x,j}^{t+1}} \tag{7}$$

being $\breve{q}_0 = \frac{D_{0:t}^t + D_{0:t}^{t+1}}{B^t}$ and $\breve{m}_0 = \frac{D_{0:t}^t + D_{0:t}^{t+1}}{B^t - D_{0:t}^{t,t+1} + \sum_{j=1}^{D_{0:t}^t} b_{0,j}^t + \sum_{j=1}^{D_{0:t}^{t+1}} b_{0,j}^{t+1}}$ the corresponding estimators for age zero.

Note that this time to calculate the total time exposed to the risk of dying in (7), (i) we compute BC as $C_{x-1}^t - D_{x-1:t-x}^t$, (ii) we subtract those dying in BCF and CFG, $D_{x:t-x}^{t,t+1}$, from BC and (iii) we aggregate the time lived with completed age x of those dying in BCF, $\sum_{j=1}^{D_{x:t-x}^t} b_{x,j}^t$, and of those dying inside CFG, $\sum_{j=1}^{D_{x:t-x}^{t+1}} b_{x,j}^{t+1}$.

In the same way as in the previous scenarios, equations (6) and (7) are also related. This time, however, the link expression is slightly more complex. The average number of years lived during their last year of life for those individuals dying age x, last birthday, f_x, is no longer constant by hypothesis, but it depends on the internal distribution of deaths at age x, and is given by expression (8):

$$\breve{f}_x = \frac{\sum_{j=1}^{D_{x:t-x}^t} b_{x,j}^t + \sum_{j=1}^{D_{x:t-x}^{t+1}} b_{x,j}^{t+1}}{D_{x:t-x}^t + D_{x:t-x}^{t+1}} \tag{8}$$

from where the exact relationship $\breve{m}_x = \frac{\breve{q}_x}{1-(1-\breve{f}_x)\breve{q}_x}$ follows.

It should be noted that the same estimator for q_x would have been obtained in the case that only aggregated counts of deaths by age, year of death and cohort had

been available. The differences would be in the estimators for m_x and f_x, which under the usual assumption of uniform distribution of deaths within each triangle would be (9) and (10), respectively:

$$\breve{\ddot{m}}_x = \frac{D^t_{x:t-x} + D^{t+1}_{x:t-x}}{C^t_{x-1} - D^t_{x-1:t-x} - \frac{2}{3}D^t_{x:t-x} - \frac{1}{3}D^{t+1}_{x:t-x}} \tag{9}$$

$$\breve{\ddot{f}}_x = \frac{\frac{1}{3}D^t_{x:t-x} + \frac{2}{3}D^{t+1}_{x:t-x}}{D^t_{x:t-x} + D^{t+1}_{x:t-x}} \tag{10}$$

2.4. Open demographic system with no hypotheses about distribution of deaths and migrants

The most data-detailed scenario considered is that in which both microdata of deaths and migrants are available. In this case, crosses, squares and circles can be accurately placed in the Lexis diagram and the amount of time that each individual is at risk of dying as a member of the target population exactly computed, the calculus of the populations exposed at risk becoming consequently more precise and the corresponding estimators more accurate. In this scenario, computing the number of person-years exposed to risk, the death probability and rate estimators are given for $\mathrm{x} = 0,1,2 \ldots, \omega,$ respectively, by:

$$\ddot{q}_x = \frac{D^t_{x:t-x} + D^{t+1}_{x:t-x}}{\ell^t_x - E^{t,t+1}_{x:t-x} + \sum_{j=1}^{E^t_{x:t-x}} e^t_{x,j} + \sum_{j=1}^{E^{t+1}_{x:t-x}} e^{t+1}_{x,j} + \sum_{j=1}^{I^t_{x:t-x}} i^t_{x,j} + \sum_{j=1}^{I^{t+1}_{x:t-x}} \mathrm{i}^{t+1}_{x,j}} \tag{11}$$

$$\ddot{m}_x = \frac{D^t_{x:t-x} + D^{t+1}_{x:t-x}}{R^t_x - D^{t,t+1}_{x:t-x} + \sum_{i=1}^{D^t_{x:t-x}} b^t_{x,i} + \sum_{i=1}^{D^{t+1}_{x:t-x}} b^{t+1}_{x,i}} \tag{12}$$

where $R^t_x = \ell^t_x - E^{t,t+1}_{x:t-x} + \sum_{j=1}^{E^t_{x:t-x}} e^t_{x,j} + \sum_{j=1}^{E^{t+1}_{x:t-x}} e^{t+1}_{x,j} + \sum_{j=1}^{I^t_{x:t-x}} i^t_{x,j} + \sum_{j=1}^{I^{t+1}_{x:t-x}} i^{t+1}_{x,j}$, $\ell^t_0 = B^t$ for age zero, $\ell^t_x = C^t_{x-1} - D^t_{x-1:t-x} + N^t_{x-1:t-x}$ for $x \neq 0$ and, again, the previous relationship $\ddot{m}_x = \frac{\ddot{q}_x}{1-(1-\breve{f}_x)\ddot{q}_x}$ applies.

Thanks to the availability of microdata, the hypotheses of closed demographic system and uniform distributions of deaths and migrants have become unnecessary. The estimators (11) and (12) indeed look assumption-free. The truth however is different. Actually a couple of implicit hypotheses have been surreptitiously introduced in their construction, as we did when the estimators (4) and (5) were defined. In particular, when constructing estimators (4), (5), (11) and (12) we have implicitly assumed that (i)

immigrants acquire the same risk of death as resident population and that (ii) those individuals who emigrate have a risk of death similar to those of the population remaining. Two assumptions, however, that could be questioned. Someone might conjecture that the probability of being a migrant should be higher for healthy people and that, therefore, some kind of selection effect may apply for migrants. Even more individual detailed statistics would be necessary to take this into account.

3. Assessing the hypotheses

In the process of developing more assumption-free estimators, two broadly used hypotheses have been questioned: the hypotheses of uniform distributions (of deaths and migrants) and the hypothesis of closed demographic systems. In this section, we evaluate their suitability in some real Spanish data-sets and assess the compatibility of the data with the expected outcomes that derived from the hypotheses. The next section is devoted to the comparison of the life tables obtained using the different estimators.

3.1. Data

Spanish death and migrant microdata by age and gender and (1 January) Population Now Cast estimates by age and gender for the years 2006–2008 constitute the data analyzed. During these years, Spain lived heavy foreign migration flows, so the database represents an interesting instance for testing the suitability of the uniform and closed demographic system hypotheses.[22]

Migrant microdata come from the Statistics of Residential Variation, which compiles, among other issues, movements from or to foreign countries by gender, including the dates of birth and of migration of each migrant. Death microdata include nationality, gender and dates of birth and death of each deceased. Population Now Cast estimates is a synthetic statistic developed by the Spanish Official Statistical Agency (INE) intended to determine at any given time the profile of the resident population in Spain, broken down by sex and age.

[22] Death Spanish microdata were provided by INE under request and on payment, migrant microdata were downloaded from http://www.ine.es/prodyser/micro_varires.htm, and Spanish Population Now-Cast estimates came from http://www.ine.es/jaxiBD/menu.do?L=1&divi=EPOB&his=0&type=db.

Migration data have a long history of being problematic and inconsistent (Kelly, 1987; Wiśniowski *et al.*, 2016, forthcoming), mostly when intense flows of irregular immigrants are recorded. Hence, it is appropriate to add some comments about the quality of Spanish migration data. Statistics of Residential Variation come from municipality registers. They are administrative registers whose record, formation, maintenance, revision and custody correspond to town councils. They are updated monthly and nationally matched by INE agents. Hence, at least in principle, a person should be included in only one register at a time. All people resident in Spain, even irregular migrants, have an incentive to report their presence in a particular municipality, as access to the health system (which was universal during the analysed period) is linked to being registered. Therefore, theoretically, all residents are included in the municipal registers.

The registers however do not necessarily reflect the exact truth. We were warned by the Office of Statistics (INE) that when an immigrant does not know the exact date of her/his birthday, this date is administratively set to the first day of January. A practice that provokes (for each gender, age and year) an obvious and artificial excess in the number of immigrants recorded as born on 1 January. Hence, before using immigrants' microdata, for each group of first-of-January-born immigrants, a number of immigrants equivalent to the artificial excess (measured as the difference over the corresponding daily mean of immigrants of the year) were randomly selected and they were randomly assigned a day of birth. These amended data-sets were the ones used to run the tests and estimate the life tables.

Likewise, as a practical check of the coherence of the data and of the impact of errors in migration flows, we have estimated BC (see Figure 1-left) in two different ways (as $\mathrm{BC} = \mathrm{C}^{t}_{x-1} - \mathrm{D}^{t}_{x-1:t-x} + \mathrm{N}^{t}_{x-1:t-x}$ and also as $\mathrm{BC} = \mathrm{C}^{t+1}_{x} + \mathrm{D}^{t}_{x:t-x} - \mathrm{N}^{t}_{x:t-x}$) and computed the discrepancies. The average of the percentage of the relative discrepancies in absolute value do not surpass in any case (for the age range 1-95) 0.20%.[23] As we

[23] For males the exact values are 0.13% in both 2006 and 2007; and for females they are 0.20% and 0.14% in, respectively, 2006 and 2007. From age 96, the average discrepancies soar, reaching 5.51%, 3.67%, 4.57% and 3.26%, respectively.

discuss when we compare life tables (see section 4), this has almost no impact on our results (see Figure 52A in the supplementary material).

3.2. Statistical tests

Numerous testable consequences can be stated from the uniform hypotheses. On the one hand, we can focus on the representation of the data on the Lexis scheme and observe the set of points where crosses, circles and squares occur as realizations of bivariate point pattern processes in the space. Under this new perspective, tools from spatial statistics can be used in order to check the suitability of the uniform distribution assumptions, whose equivalent in spatial terms are the hypotheses of complete spatial randomness (CSR)[24]. Hypotheses in our case must be tested from a period, cohort and period-cohort point-of-view in, respectively, Lexis (1x1)-cells (squares), Lexis quadrilaterals (e.g., BCEF in Figure 1-left) and Lexis triangles.

Complete spatial randomness of death and migrant events in Lexis cells, triangles and quadrilaterals has been evaluated running, in version 3.0.2 of the statistical software R (R Core Team, 2013), three of the tests available in the R package spatstat (Baddeley and Turner, 2005). All the three tests implemented have CSR as null hypothesis, varying in their decision rules and/or in their alternative hypotheses. Specifically, the implemented tests have been the Maximum Absolute Deviation (MAD) test (Ripley 1977, 1981), the CLF test (Cressie, 1991; Loosmore and Ford, 2006) and a spatial version of the well-known chi-squared goodness-of-fit (XS) test (see, e.g. Grabarnik and Chiu, 2002). Both MAD and CLF tests are based on comparing a summary statistic of the deviations of the observed pattern across the distances of interest with the corresponding distribution of deviations of an envelope of simulated point patterns (in the same space and with the same number of points as our data) from a spatial Poisson process. The MAD test uses the MAD distance as summary statistic whereas the CLF uses the total squared deviation (Baddeley *et al.*, 2014). To perform the XS test, triangles have been divided into eight polygons of equal size and Lexis cells (squares) and quadrilaterals into sixteen equal-size polygons and the number of points in each

[24] According to the probability Laplace principle, the CSR assumption entails that the number of point events located in any arbitrary sub-region *A* of a target region *D* is proportional to its area.

polygon compared (using the Pearson chi-squared distance) with the corresponding expected number of points in each polygon under uniform distribution.

On the other hand, under a probability distribution perspective, uniform hypotheses also entail testable outcomes. Thus, for example, under the death uniform assumption, the variable measuring the years lived (or exposed to risk of dying) in their last year of life (or in a given calendar year *t*) for those individuals dying with completed age x would be distributed as a uniform random variable in the interval $[0,1)$. Furthermore, the same would be true for the distributions of the time exposed to risk of dying of emigrants and immigrants. What's more, the density function of the random variable, τ, measuring the time exposed to risk at any age x of either deaths or migrants would be (a) $f(\tau) = 2 - 2\tau$, for $0 \leq \tau < 1$, when deaths and emigrants are located in lower triangles or when immigrants are located in upper triangles, and (b) $f(\tau) = 2\tau$, for $0 \leq \tau < 1$, in the mirror case.[25] To test these functional distribution hypotheses, we have used the well-known Kolgomorov-Smirnov (KS) test (see, e.g. Li and Racine, 2007) applying the R function ks.test, as well as the geometric (G) test available in the R package GoFKernel (Pavía, 2015). The G test is based on measuring the discrepancy on the L1-norm between a kernel density function estimate of the observed data and the null hypothesis density function.

In the estimators proposed in the previous section, however, the uniform assumptions materialize in precise outcomes, much more specific than the distributions in the above paragraph. Hence, in addition to testing the acceptability of the general uniform assumptions, we also examine whether: (i) the number of death events occurring in upper and lower triangles of the same age and period are equal (employed

[25] Using the well-known result that the density function of a uniform bivariate variable in a surface S of area A is equal to $f(t_1, t_2) = \frac{1}{A}$ for $(t_1, t_2) \in S$, a way to arrive at the above distributions is as follows. Taking B in Figure 1 as the origin of a Cartesian coordinate system, we have under the hypothesis of uniform distribution of deaths by age and calendar year that the density distribution functions of an event occurring respectively in the whole square, the lower triangle and the upper triangle can be expressed by: $f_C(t_1, t_2) = 1$ for $0 \leq t_1, t_2 < 1$, $f_I(t_1, t_2) = 2$ for $0 \leq t_2 \leq t_1 < 1$ and $f_S(t_1, t_2) = 2$ for $0 \leq t_1 \leq t_2 < 1$. From which, observing that the years lived for those individuals dying aged x last birthday in their last year of life is equal to t_2, we have after calculating the marginal distributions of t_2 the densities of interest: $f_C(t_2) = \int_0^1 d\, t_1 = 1$, $f_I(t_2) = \int_{t_2}^1 2\, d\, t_1 = 2t_1]_{t_2}^1 = 2 - 2t_2$ and $f_S(t_2) = \int_0^{t_2} 2\, d\, t_1 = 2t_1]_0^{t_2} = 2t_2$. The distributions of the time exposed to risk of dying for emigrants and immigrants are obtained in a similar fashion.

in equations (1) to (5)); (ii) the number of immigrant and emigrant events in upper and lower triangles (of the same age and period) are equal (used in equations (4) and (5)); (iii) the average time exposed to risk at age x is $\frac{1}{3}$ for deaths located in lower triangles and $\frac{2}{3}$ for deaths located in upper triangles (applied in equations (2), (3), (5), (9) and (10)); (iv) the average time exposed to risk at completed age x is $\frac{1}{3}$ $\left(\frac{2}{3}\right)$ for emigrants (immigrants) located in lower (upper) triangles and $\frac{2}{3}$ $\left(\frac{1}{3}\right)$ for emigrants (immigrants) located in upper (lower) triangles (utilized in equations (4) and (5)); and, (v) the average number of years lived for those individuals dying aged x, last birthday in their last year of life is $\frac{1}{2}$ (employed in equation (3)). Two-side binomial tests with $p = 0.5$ have been used to test the particular hypotheses (i) and (ii) and two-side mean t-student tests to assess hypotheses (iii) - (v).

Finally, we study the suitability of the hypothesis of closed demographic system. In addition to comparing in the next section the life tables constructed with and without migratory flows, we assess in Subsection 3.6 two of the three implicit assumptions on which the practical feasibility of this hypothesis rests. We have performed several tests to evaluate the assumptions of (i) considering that entries and exits by age and gender are not significant and (ii) admitting that, for each age and gender, migration flows are random and show closely similar input and output figures. The third implicit assumption consisting in (iii) supposing that migrants share the same risk of death as resident population is untestable with our data. Appraising (ii) does not pose special problems, two-side binomial tests with $p = 0.5$ can be employed in each period and cohort age–gender group. Testing (i), however, presents more difficulties. The key point here is to determine when a number of immigrants (emigrants or net migrants) is significant, i.e. when it could be considered large enough, compared to the size of the target population, as to challenge the closed demographic system hypothesis. To determine this threshold, we take as reference the definition of rare disease in Europe – which states that a disease or disorder is rare when it affects less than 1 in 2000 inhabitants (see, e.g. Posada *et al.*, 2008) – and, following a conservative approach, we consider that a rate of immigrants (or net migrants, in absolute value) or a probability of emigrating[26]

[26] In the case of immigrants we cannot talk of probability given that before immigration they are not members of the target population. Emigrants, on the other hand, are exits from the target population.

is significant when it is statistically larger than ten times the chance of developing a rare disease, i.e. 0.5%. In particular, to ascertain whether migrant figures are significant for each gender, age and calendar year (Lexis cells) and for each gender, age and cohort (quadrilaterals), we have performed one-side binomial tests with $p = 0.005$.

A total of 43,206 hypothesis tests have been completed.[27] Hence, even with all null hypotheses being correct, we would expect, at the usual significant levels 10, 5 and 1%, approximately 4321, 2160 and 432 rejections occurring, respectively, by chance. Thus, instead of reporting all rejections, we will focus on those outcomes that show consistency among tests, genders and periods. To avoid increasing the length of the article by a huge amount, when assessing the hypotheses of uniformity of death and migrant events (Subsections 3.3 and 3.4) we will just present in the main text those tests corresponding to the 2006–2007 quadrilaterals and the 2007 cells and triangles. The outputs of all these tests are available in the supplementary material.

3.3. Death uniformity

As a rule the hypothesis of uniform distribution of deaths by age and calendar year is extensively accepted. According to our outcomes, this assumption could be employed with confidence for the range of ages for which the force of mortality is not too strong. However, as soon as the probability of death grows the use of detailed data should be promoted when available.

As expected, the hypothesis of uniform distribution of deaths is systematically rejected for age zero. Hence, as it is broadly acknowledged, special expressions must be employed with aggregated data for q_0 and m_0. Furthermore, as can be inferred looking at Figure 3 —where 2007 graphical summaries of the outcomes of the death uniform tests performed at the usual significant levels (0.1, 0.05 and 0.01) are displayed for different Lexis surfaces— the uniform assumption for deaths tends to be mistaken from

[27] We have considered data of three calendar years (2006–2008), have dealt independently with males and females and have studied the range of ages from 0 to 108 (yet some tests could have been performed up to age 112). In particular, we have completed 19,008 spatial tests, 14,388 functional tests, 6322 point parametric tests and 3488 tests to assess the hypothesis of closed demographic system.

about the age of 65 (the Spanish age of legal labour retirement at the time).[28] After age 65, there are a large number of cases for which the hypothesis of uniform distribution of deaths looks inappropriate.

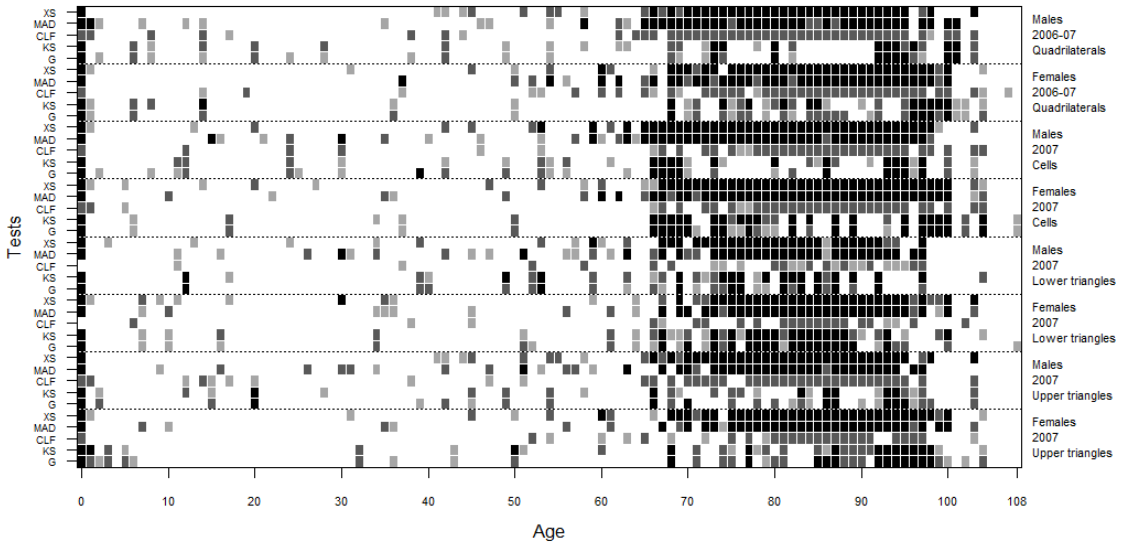

Figure 3. Death uniform hypothesis tests by gender and age in Spanish population for people dying in 2006-07 Lexis quadrilaterals, 2007 Lexis cells (squares), 2007 Lexis lower triangles and 2007 Lexis upper triangles. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. MAD denotes the Maximum Absolute Deviation test, CLF the Cressie-Loosmore-Ford test, XS the spatial Chi-squared goodness-of-fit test, KS the Kolgomorov-Smirnov test and G the Geometric test. The three first tests are spatial tests and check complete spatial randomness as null hypothesis after a representation of events in the Lexis space. KS and G tests are nonparametric tests and check whether sample exposed-to-risk times are compatible with the assumed probability distributions.

3.4. Migrant uniformity

In order to assess the hypotheses of uniform distribution of migrants in our data-set, a large battery of tests has been completed. Figures 4 and 5 present for emigrant and immigrant events graphical summaries of the results of the tests performed at the usual significant levels (0.1, 0.05 and 0.01) for different Lexis surfaces focused on 2007.[29] The results of the tests clearly reveal a failure of these assumptions for a large range of ages.

[28] The same picture is obtained scrutinizing in the supplementary material the Figures 1A to 4A. Figures 1A to 4A display the outputs of the tests for the whole database.

[29] Figures 5A to 12A in the supplementary material offer the results for the whole period covered by our database. These figures display the same overall picture.

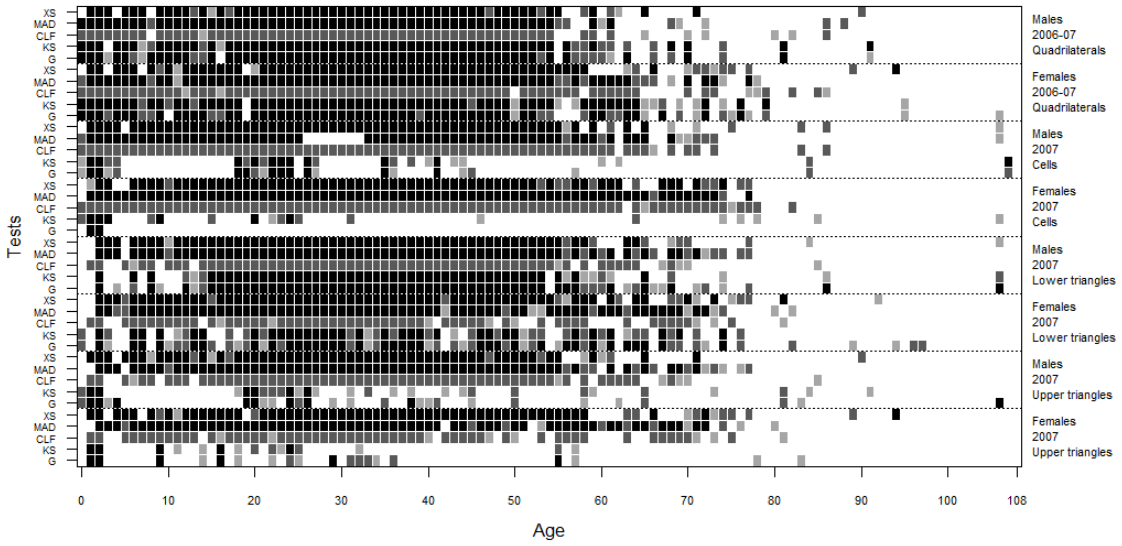

Figure 4. Uniform hypothesis tests by gender and age in Spanish population for emigrant events occurring in 2006-2007 Lexis quadrilaterals, 2007 Lexis cells (squares), 2007 Lexis lower triangles and 2007 Lexis upper triangles. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. Details of the tests can be found in the text. MAD denotes the Maximum Absolute Deviation test, CLF the Cressie-Loosmore-Ford test, XS the spatial Chi-squared goodness-of-fit test, KS the Kolgomorov-Smirnov test and G the Geometric test. The three first tests are spatial tests and check complete spatial randomness as null hypothesis after a representation of events in the Lexis space. KS and G tests are nonparametric tests and check whether sample exposed-to-risk times are compatible with the assumed probability distributions.

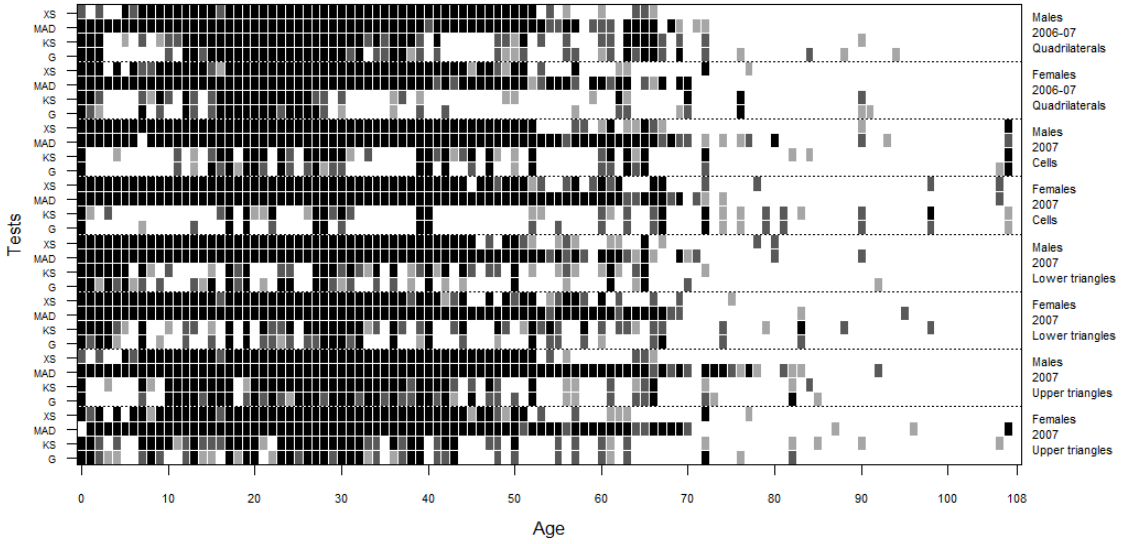

Figure 5. Uniform hypothesis tests by gender and age in Spanish population for immigrant events occurring in 2006-2007 Lexis quadrilaterals, 2007 Lexis cells (squares), 2007 Lexis lower triangles and 2007 Lexis upper triangles. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. MAD denotes the Maximum Absolute Deviation test, XS the spatial Chi-squared goodness-of-fit test, KS the Kolgomorov-Smirnov test and G the Geometric test. CLF tests were not performed in this case because the R function dclf.test was unable to handle the large number of immigrant events occurring in the majority of surfaces. The two first tests are spatial tests and check complete spatial randomness as null hypothesis after a representation of events in the Lexis space. KS and G tests are nonparametric tests and check whether sample exposed-to-risk times are compatible with the assumed probability distributions.

In light of the available outcomes, we can asseverate that, despite the logic behind the hypotheses of uniform distribution of migration events, migration flows were not distributed randomly in Spain during the period ranging from 2006 to 2008. Indeed, a deep analysis of the data points to some kind of seasonality in migration flows, mainly in emigration events. Overall, the averages time exposed to risk of migrants show figures systematically above the theoretical ones[30].

3.5. Particular hypotheses

The results of the tests performed to evaluate the hypotheses of uniform distributions of deaths and migrants lead us to consider these assumptions as inadequate in the whole range of ages and therefore point to the use of detailed data when available. The above conclusions are strongly reinforced when we analyse the expected consequences of the hypotheses and consider the suitability of their materializations in the formulae currently used to estimate raw mortality rates.

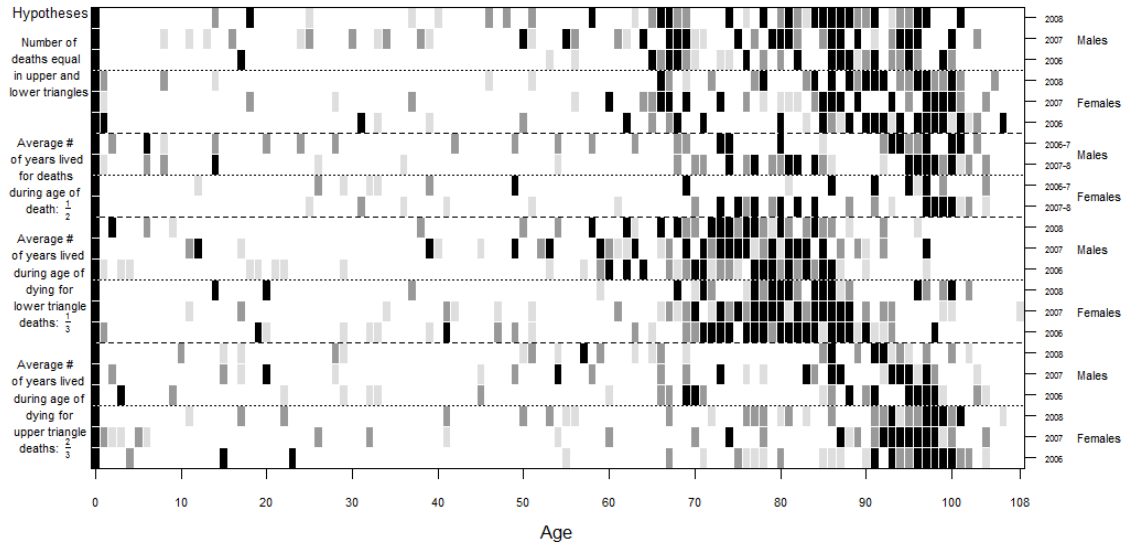

Figure 6. Parametric hypothesis tests by gender and age for 2006 to 2008 Spanish population, corresponding to the concreteness of the hypotheses of uniform distribution of deaths in equations (1) to (5), (9) and (10). Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. Two-side binomial tests with $p = 0.5$ are used to test the first block of hypotheses and two-side mean t-student tests to assess the next three blocks of hypotheses.

In particular, if the concreteness of the hypotheses of uniform distribution of deaths are systematically inadequate from age 60 (see Figure 6), the concreteness of the hypotheses of uniform distribution of migrants are clearly disappointing up to

[30] This can be observed in Figures 21A to 31A in the supplementary material.

approximately age 70 (see Figures 7 and 8). As can be observed in Figure 6, the hypothesis of equal number of deaths occurring in lower and upper triangles per age and calendar year are frequently rejected in the range of ages from 65 to 100 (53% of the times with a significant level, α, of 0.05, and 65% of the times when $\alpha = 0.01$) and the same happens an significant number of times for the hypothesis of $\frac{1}{2}$ for the average of time lived with age x for those dying before reaching age $x + 1$ (23% of times the hypothesis is rejected for $\alpha = 0.01$ in the age range 65-100). Likewise, the hypothesis of similar number of migrants in upper and lower triangles of the same Lexis cell is also recurrently rejected up to age 60 (for example, with $\alpha = 0.01$ it is rejected 93% of the times for emigrants and 76% for immigrants) and the same happens regarding the null hypothesis about the averages of time exposed to the risk of dying in the target population for those immigrating and emigrating (Figures 7 and 8). On average and up to age 60, these hypotheses are rejected 56% of the times for $\alpha = 0.01$.

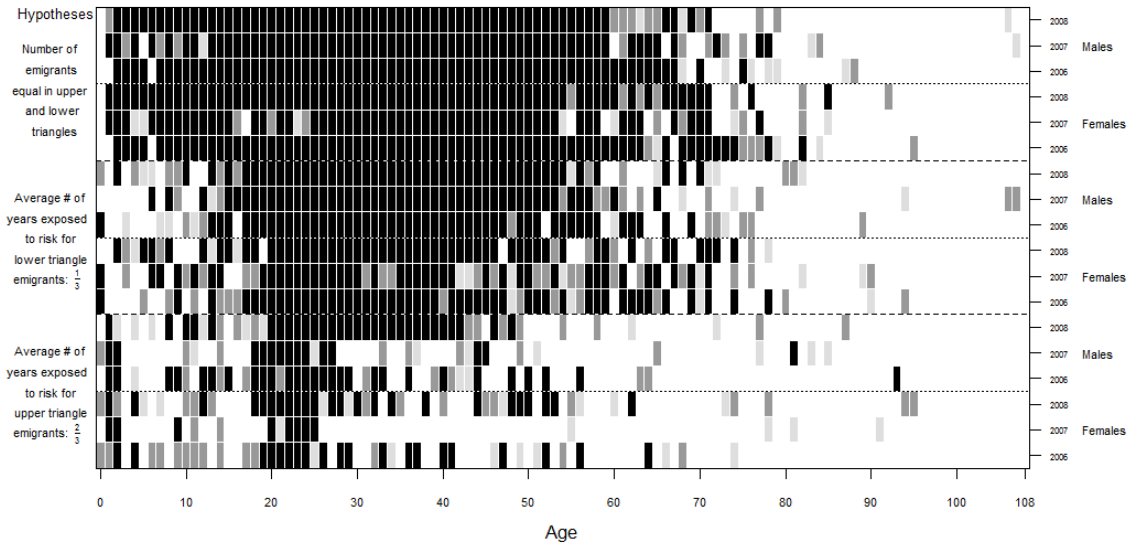

Figure 7. Parametric hypothesis tests, by gender and age for 2006-2008 Spanish population, corresponding to the concreteness in equations (4) and (5) of the hypotheses of uniform distribution of emigrants. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. Two-side binomial tests with $p = 0.5$ are used to test the first block of hypotheses and two-side mean t-student tests to assess the next two blocks of hypotheses.

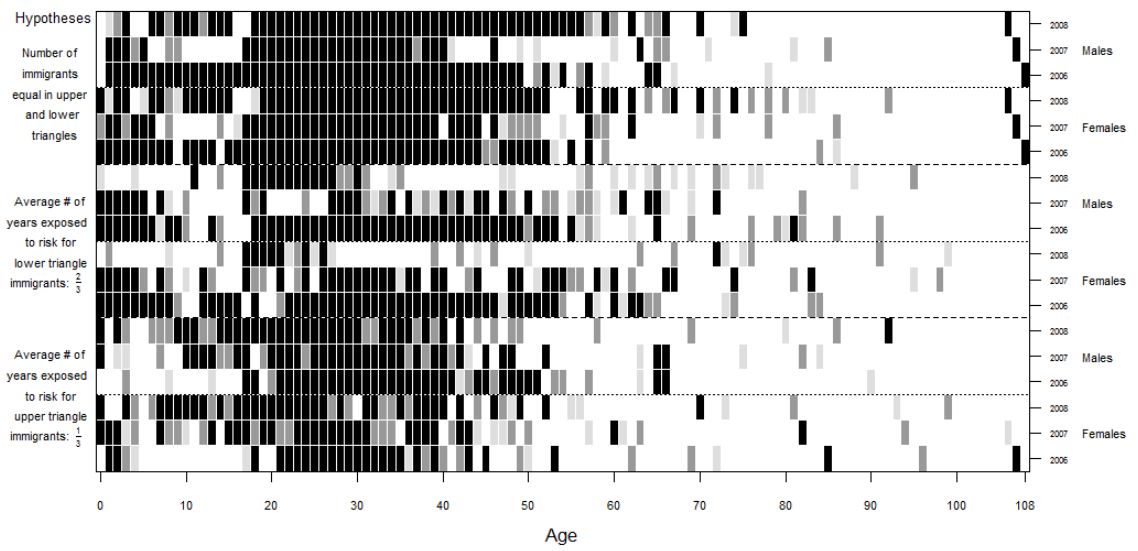

Figure 8. Parametric hypothesis tests, by gender and age for 2006-2008 Spanish population, corresponding to the concreteness in equations (4) and (5) of the hypotheses of uniform distribution of immigrants. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. Two-side binomial tests with $p = 0.5$ are used to test the first block of hypotheses and two-side mean t-student tests to assess the next two blocks of hypotheses.

3.6. Closed demographic system

We have shown that, in our data-set, the hypotheses of uniform distribution of emigrants and immigrants and, mainly, their particular consequences are inappropriate. In spite of this, it could be stated that their impact was insignificant. In other words, that the relative weights of migration flows were so scarce (compared to the sizes of the target population) as to not worry about them. Although this may be the case in other historical periods or territories, this does not seem to be true in the case of Spain for the years collected in our database. As the results displayed in Figures 9 and 10 clearly show, the flows of emigrants and mainly of immigrants (which were during the studied period significantly more common in Spain) were really significant and should be considered explicitly.

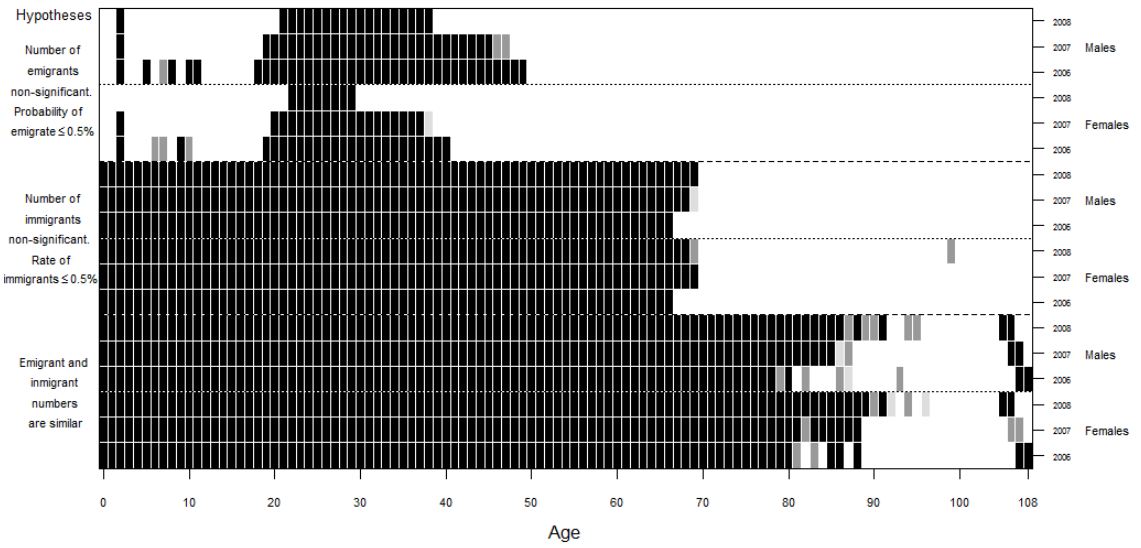

Figure 9. Some closed demographic system hypothesis tests by gender and age for 2006-2008 Spanish population. Tests of whether entries and exits can be thought of as relatively rare events and whether input and output migration figures compensate each other are displayed. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. One-side binomial tests with $p = 0.005$ are used to test the two first blocks of hypotheses and two-side binomial tests with $p = 0.5$ to assess the third block of hypotheses.

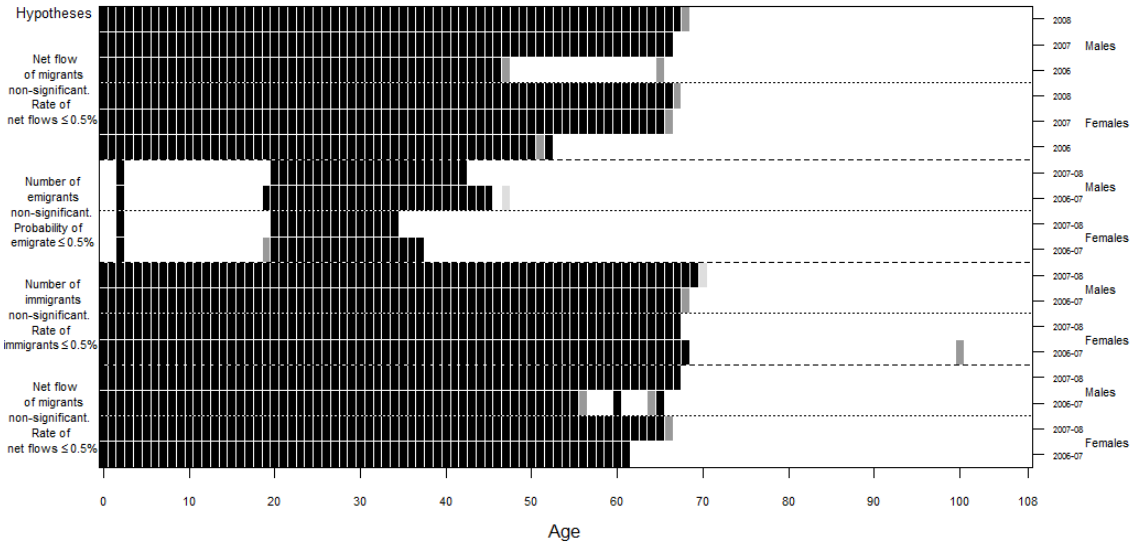

Figure 10. Additional closed demographic system hypothesis tests by gender and age or cohort for 2006-2008 Spanish population. Tests of whether entries, exits and net migrant flows can be thought of as relatively rare events displayed. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. One-side binomial tests with $p = 0.005$ are used to evaluate the hypotheses.

Both in cohort and in year terms, the number of immigrants that choose Spain as destination represented a significant amount of the Spanish resident population up to even the age of 70 years. What's more, emigrant flows, which were well significantly below immigrant flows, are only significant for the range of ages from 20 to

approximately 40, the range of ages of outflows of males being, nevertheless, wider than the range of women.

4. Comparing life tables

In the previous section, a huge number of statistical tests have been performed to evaluate the suitability of the hypotheses of uniform distributions of deaths and migrants and of closed demographic system. In the light of the outcomes, these hypotheses do not look appropriate, at least for the database considered, and consequently more free-assumption estimators should be encouraged from a theoretical perspective. In this section we study the impact of the hypotheses in practice, on the estimated death probabilities. In particular, we compare the life tables obtained using the different estimators and, when relevant, we analyse the deviations observed within the same table between the direct estimates of q_x and the indirect approximations of q_x reached via m_x.

Life tables are usually constructed separately for men and for women because of their very different mortality patterns. Consequently, a total of sixteen life tables have been constructed after combining the two genders (men and women) and the four scenarios (closed demographic system and uniform distribution of deaths by age and calendar year, CDS_UD; open demographic system and uniform distribution for deaths and migrants, ODS_UDM; closed demographic system with no hypothesis about distribution of deaths, CDS_NH; and, open demographic system with no hypotheses about distribution of deaths and migrants, ODS_NH) with two cohort periods (2006–2007 and 2007–2008). As is well known, however, life tables of crude mortality probability estimates are subject to random fluctuations, an issue that can mask or exacerbate some of the differences. Thus, as is a common practice in demography and actuarial sciences, the crude estimated life tables have been graduated (adjusted) in order to smooth the profile of the associated stochastic processes. The smoothing has been carried out using nonparametric Gaussian kernel graduation (see, e.g. Ayuso *et al*., 2007, pp. 217–222) with a window parameter, or bandwidth, equal to 2. Both crude and graduated life tables have been used in comparisons.

The differences between the probabilities estimated under each scenario are scarce in absolute values.[31] This is a consequence of the fact that death probabilities are (fortunately) small for almost all ages. Hence, the comparisons between the values obtained for each age x in the different scenarios have been performed using as dissimilarity indicator a relative measure; in particular, the absolute relative discrepancy. All the possible comparisons between scenarios, however, do not make sense. Because the CDS_UD scenario can be viewed as a particular instance of the ODS_UDM, CDS_NH and ODS_NH scenarios and the ODS_UDM and CDS_NH scenarios as simpler examples of the ODS_NH scenario, we restrict ourselves to comparing the death probabilities of the nested scenarios. In the computation of the absolute relative discrepancy measures, the probabilities belonging to the less detailed-data demanding scenario are always taken as reference. Furthermore, given that the great differences among estimated probabilities are concentrated in the oldest ages (more than 100 years old), characterized by having much smaller exposed to risk populations and less reliable data, the comparisons have been constrained to the range of ages from 0 to 100 years to avoid that these numbers dominate the scrutiny.

Regarding comparisons between direct q_x and indirect q_x (via m_x) estimates[32], we see that in general the differences are quite insignificant; yielding both estimation approaches at the end to almost the same life tables. It is worth mentioning, nevertheless, that as expected under the CDS_UD scenario some differences start to appear as soon as the null hypothesis of uniform distribution of deaths is challenged by the data. The small differences observed between both ways of estimating crude mortality rates in the ODS_UDM scenario show a less clear pattern, the latter being a mixture of the failings of the uniform hypotheses for deaths and migrants.

If the differences between estimating q_x directly or via m_x within a given scenario are negligible, the same cannot be said when we compare the life tables reached under each scenario, as the graphical summaries of the comparatives of the

[31] This can be observed in Figures 37A and 38A and Figures 39A and 40A available in the supplementary graphical appendix, where respectively the crude and graduated estimated life tables have been displayed in the log-scale.

[32] Graphically available in Figures 41A and 42A in the supplementary material.

absolute relative discrepancies by age between the nested scenarios (see Figure 11 and also in Figures 43A to 51A in the supplementary material) show.

In what refers to comparisons between the uniform-free scenarios (CDS_NH and ODS_NH) versus the uniform-based scenarios (CDS_UD and ODS_UDM), we notice that as a rule the differences for crude probabilities are higher than those for graduated probabilities.[33] Moreover, despite the mountain-peaking shape that the discrepancies show, two patterns can be discerned. We observe that on average the differences declined with age and, what looks more interesting, the shapes between tables of the same period are closer than those of the same gender. The latter is likely revealing that calendar-risk-related issues (such as, the strength of the epidemic influenza of the year or the seasonality of migrant flows) are more important than gender-risk-related issues when referring to the failure of the uniform hypotheses. Finally, within these comparisons, it is worth mentioning the peaks that can be observed (mainly for non-graduated rates) corresponding to q_{66} and q_{67} and to q_{67} and q_{68} for the tables of 2006-07 and 2007-08, respectively. This issue suggests that the number of births during 1940 (the year just after the end of the Spanish Civil War) was significantly higher than the number of births registered in both 1939 and 1941.

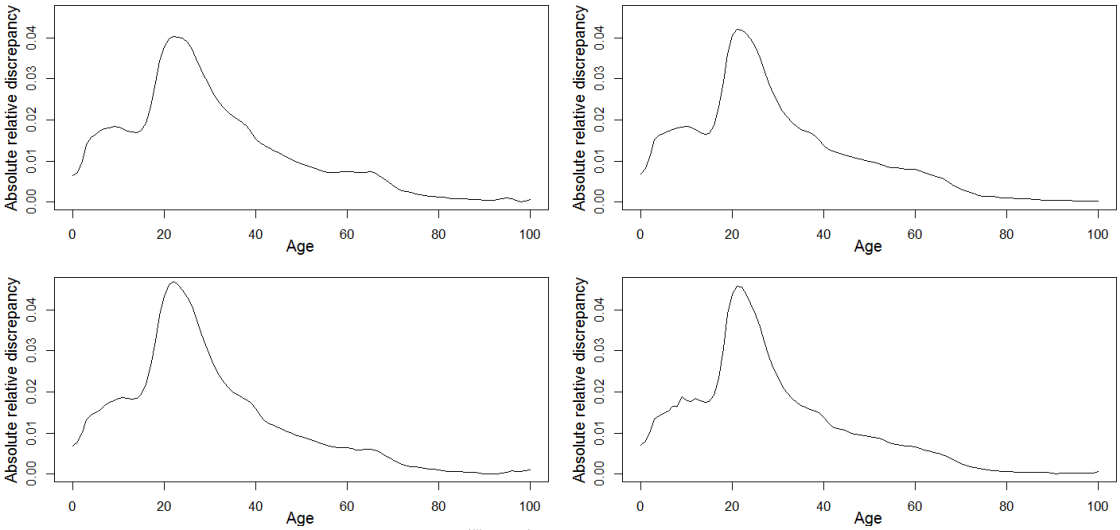

Figure 11. Absolute relative discrepancies, $\frac{|\ddot{q}_x - \bar{q}_x|}{\bar{q}_x}$, between the graduated probabilities of death obtained under the closed demographic system with no hypothesis about distribution of deaths (CDS_NH) scenario and the open demographic system with no hypotheses about distribution of deaths and migrants

[33] This can be observed comparing Figure 45A versus Figure 50A and Figure 47A versus Figure 52A in the supplementary goerimaterial.

(ODS_NH) scenario for 2006-2007 men (upper left), 2006-2007 women (upper right), 2007-2008 men (lower left) and 2007-2008 women (lower right).

Focusing now on the comparisons between closed and open demographic system scenarios (CDS_ UD vs. ODS_UDM and CDS_NH vs. ODS_NH), one can deduce that, in contrast to the previous comparisons, the discrepancies between graduated probabilities and crude probabilities are of the same order of magnitude [34] and that, in our database, they can surpass even 4% of discrepancy. As expected, the patterns of the differences mimic that of migration flows and mainly those of immigrant flows because immigration flows were well above emigration flows during those years in Spain. Likewise, despite the shapes of the differences being quite similar this time between periods, these contours are again more alike between genders of the same period than between contours of the same gender in different periods.

Continuing with the analysis of the impact of migration flows and given the well-known long history of lack of quality of migration statistics, it is worth asking ourselves about the impact of this issue on our estimates. In our theoretical exposition, we have chosen to estimate the exposed-to-risk population at age *x*, BC, using as a base AB but the same can be computed from FC (see Figure 1). With perfect data, both ways would yield exactly the same results. We have already seen that the differences between obtaining BC from AB and from FC are quite small (see subsection 3.1). Now, we briefly examine, focusing on those ages for which migration flows are significant, the impact of these discrepancies in the graduated probabilities. In light of the outcomes, our results are quite robust. There are almost no differences in the estimated rates. The discrepancies are of an order of magnitude ten times smaller than the ones attained comparing CDS_NH and ODS_NH probabilities. For example, in the ODS_NH scenario, the average of the absolute relative discrepancies in the graduated probabilities is just 0.12% and never surpass 0.5% for the range of ages 1-80 (see Figure 52A).

Comparing the least data demanding scenario (CDS_UD) vs. the most detailed-data scenario (ODS_NH), we observe that, for the instances handled in this paper and as a rule, the discrepancies caused by the break of uniform hypotheses dominate over the

[34] See Figures 43A and 48A, for raw tables, and Figures 49A and 12 for graduated tables.

discrepancies due to the hypothesis of closed demographic system[35], both for crude rates and for graduated probabilities. Obviously, as happened when comparing uniform versus non-uniform scenarios, the graduation step mitigates discrepancies.

Finally, we study how the misestimates of mortality attained—provoked by the failure of uniformity of deaths in some ages, the seasonal behaviour of migration events and the higher amounts of immigrants than of emigrants—translate with the less data-demanding estimators in terms of life expectancy and other financial-actuarial measures (like annuities or premiums). With the aim of making a broader comparison, we also include in the assessment the same figures derived from the Human Mortality Database (HMD) and INE life tables. Although INE and HMD tables were constructed using as building blocks period-based estimators (under the assumption of uniform distribution of deaths inside each Lexis triangle and without explicitly considering migration flows), we include them in the scrutiny because of their national and international role as referents. INE elaborates the Spanish official life tables and HMD provides estimates internationally accepted.[36]

Table 3. Examples of summary statistics from different life tables.

	Males					Females				
	q_0	e_0	e_{30}	e_{65}	ℓ_{65}	q_0	e_0	e_{30}	e_{65}	ℓ_{65}
CDS_UD	0.00374	77.77	48.71	17.75	841,903	0.00321	84.08	54.68	21.67	930,320
ODS_UDM	0.00372	77.84	48.76	17.76	843,258	0.00319	84.11	54.70	21.68	930,907
CDS_NH	0.00382	77.75	48.70	17.77	840,825	0.00330	84.09	54.70	21.71	929,665
ODS_NH	0.00379	77.82	48.75	17.78	842,279	0.00328	84.13	54.72	21.71	930,271
INE	0.00366	77.51	48.42	17.33	844,429	0.00322	83.83	54.41	21.34	931,718
HMD	0.00364	77.77	48.68	17.63	844,822	0.00321	84.15	54.73	21.68	931,803

CDS_UD: Closed demographic system and uniform distribution of deaths by age and calendar year.
ODS_UDM: Open demographic system and uniform distribution for deaths and migrants.
CDS_NH: Closed demographic system with no hypothesis about distribution of deaths.
ODS_NH: Open demographic system with no hypotheses about distribution of deaths and migrants.
INE: Instituto Nacional de Estadística.
HMD: Human Mortality Database.
ℓ_{65}: Number of survivors to exact age 65 of a fictitious initial cohort of 1,000,000 live births.

We focus on the 2007 tables and, to make comparisons fair, we work with raw probabilities, set $q_{101} = 1$ and round probabilities to five decimal places before making

[35] See Figures 45A and 50A in the supplementary material.

[36] Both tables are almost equal up to age 75, age from which the effect of the modelling strategy that HMD employs for older ages starts to be evident.

computations[37]. Table 3 offers some examples of actuarial-demography statistics calculated from the different tables. In particular, it contains information about infant mortality rates, number of survivors to exact age 65 and life expectancies at birth and for ages 30 and 65. As expected, all the figures are quite close, but some patterns emerge. Infant mortality probabilities and life expectancies are, as a rule, higher in our tables, being however the number of survivors at age 65 smaller. Focusing on our estimates, the deviations introduced by the uniform hypotheses have the global effect of an underestimation of mortality (see also Table 4). It seems that the positive net migration movements recorded in Spain in 2007 provoke a global underestimation of probabilities in closed demographic scenarios for ages up to 65 and that the hypothesis of uniform distribution of deaths causes a slight underestimation of probabilities after age 65. The results displayed in Table 4 corroborate these impressions.

Table 4 shows some examples of the differences that in terms of annuities and premiums would entail using each life table[38]. In particular, assuming a null discount rate and working with unloaded probabilities, we present the differences (measured in days) to be paid for the insurer in an immediate fixed life annuity of a annuitant of 65 years, $\ddot{a}_{65}$, and the differences (measured in percentage) of premiums to be paid for a 20 year old insured for a year term life insurance, ${}_{1}A_{20}$. As can be seen, the deviations that the hypotheses introduce in each age do not balanced out, and some differences remain depending on the table applied. For example, from using the ODS_NH table to using the CDS_UD table, we find a global average difference of almost two weeks per person. An issue of great importance given that there are more than 5,8 million of retired people currently living in Spain with a monthly average pension slightly higher than 1,000€ per beneficiary, which represents more than 2,900 million of €.

[37] The INE table offers probabilities up to age 100 and the HMD table is delivered with just five decimal places.

[38] Table 4 displays numbers differentiated by gender for the sake of comparison. Currently, and because of the Test-Achats case, gender cannot be used in the EU to discriminate premiums and benefits under insurance contracts.

Table 4. Examples of differences in annuities (in days) and premiums (in percentage) with several life tables.

Males	CDS_UD	ODS_UDM	CDS_NH	ODS_NH	INE	HMD
CDS_UD		-5.21	-6.47	-11.85	54.62	141.39
ODS_UDM	-4.69%		-1.26	-6.64	59.83	146.59
CDS_NH	3.13%	8.20%		-5.38	61.09	147.86
ODS_NH	-1.56%	3.28%	-4.55%		66.47	153.24
INE	-4.69%	0.00%	-7.58%	-3.17%		86.76
HMD	-4.69%	0.00%	-7.58%	-3.17%	0.00%	

Females	CDS_UD	ODS_UDM	CDS_NH	ODS_NH	INE	HMD
CDS_UD		-2.41	-12.93	-15.48	32.54	-5.02
ODS_UD	-4.17%		-10.51	-13.07	34.95	-2.61
CDS_NH	-4.17%	0.00%		-2.56	45.47	7.90
ODS_NH	-8.33%	-4.35%	-4.35%		48.02	10.46
INE	-8.33%	-4.35%	-4.35%	0.00%		-37.56
HMD	-8.33%	-4.35%	-4.35%	0.00%	0.00%	

CDS_UD: Closed demographic system and uniform distribution of deaths by age and calendar year.
ODS_UDM: Open demographic system and uniform distribution for deaths and migrants.
CDS_NH: Closed demographic system with no hypothesis about distribution of deaths.
ODS_NH: Open demographic system with no hypotheses about distribution of deaths and migrants.
INE: Instituto Nacional de Estadística.
HMD: Human Mortality Database.
In the upper triangles, differences in days to be paid in an immediate fixed life annuity of an annuitant of 65 years, $\ddot{a}_{65}$, when the row life table is used instead of the column life table. For example, the value 61.09 in the cell (3, 5) for males means that (in average) a company must pay 61.09 more days using the CDS_NH table than the INE table. The discount rate is assumed null.
In the lower triangles, differences in percentage between the premiums to be paid for an insured of 20 years for one year term life insurance, ${}_1A_{20}$, when the row life table is used instead of the column life table. For example, the value -8.33% in the cell (4, 1) for females means that a company may *commercialise* the corresponding one year term life insurance 8.33% cheaper if the ODS_NH table were used instead of the CDS_UD table. The discount rate is assumed null.

5. Conclusions

Up to recently, collecting, transmitting and storing detailed data was a prohibitive task when millions of microdata had to be handled. The IT revolution has made it possible. Currently, even personal computers can deal with really huge amounts of data. Despite this, in life table construction, tradition has meant that detailed demographic microdata remain unexploited and only aggregated death and census figures are used. This demands that several hypotheses must be implicitly assumed during their construction. This paper assesses the suitability of (some of) these hypotheses and shows new estimators to exploit the more abundant and reliable microdata that current statistical systems offer. In particular, we take data from Spain corresponding to years 2006 to 2008 and gauge, using thousands of hypothesis tests, the suitability of the hypotheses of uniform distribution of death and migrant events and the hypothesis of closed demographic system. Moreover, within the cohort-based family of estimators of death probabilities, we borrow from the statistical-actuarial literature and suggest some new estimators that allow the incorporation of all the available detailed data in the process

of estimating crude mortality probabilities, making it possible to obtain more efficient estimates.

In light of the results, the hypotheses of uniform distribution of deaths and migrants and of the closed demographic system are not appropriate (at least for the database analysed in this research) and consequently the proposed estimators should be encouraged from a theoretical perspective. Indeed, as our outcomes in section 4 show, important differences are found depending on the estimator employed and therefore it does matter what estimator is used. It is worth remembering the great significance that mortality figures have on current societies for policy making and public planning, for instance in pension systems.

In this paper we have dealt with (biannual-period) cohort-based estimators, although obviously it is straightforward to extend the estimators proposed in this paper to the family of period-based estimators. In this case, nevertheless, the suitability of hypothesis of uniform distribution of birthdates for each age and calendar year should also be tested to safeguard the correctness of the estimator; in other case, some additional refinement should be introduced in the above expressions to maintain their exactness. Even though there would be another implicit hypothesis, whose assessment is much more difficult, that would still remain present in the estimators of both families. These assumptions are that immigrants acquire the same risk of death as the resident population and that emigrants have the same risk of death as those who do not emigrate. To test the suitability of these hypotheses longitudinal datasets for migrants would be needed.

Although heterogeneity is present in all populations, other heterogeneities than that described by time and age are not usually considered in general population life table construction. In our opinion, however, studying migrant heterogeneity could be relevant in the same way as it is in life insurance. It is foreseeable that some kind of selection effect applies mainly for labour migrants and that, consequently, this affects their relative contribution to the exposed to risk population and, through this, to the estimated rates. Likewise, when microdata are not available an interesting topic to be studied would be to analyse if either, as we think, it would be better to make intra-age adjustments (perhaps in a Bayesian framework) for death and migrant aggregate

figures, using information from a related area and/or time-period, or, on the other hand, it would be better to construct the table just using the usual assumptions.

Supplementary material. Graphical Appendix (SA2)

Assessing Implicit Hypotheses in Life Table Construction

DEATH UNIFORM HYPOTHESIS TESTS

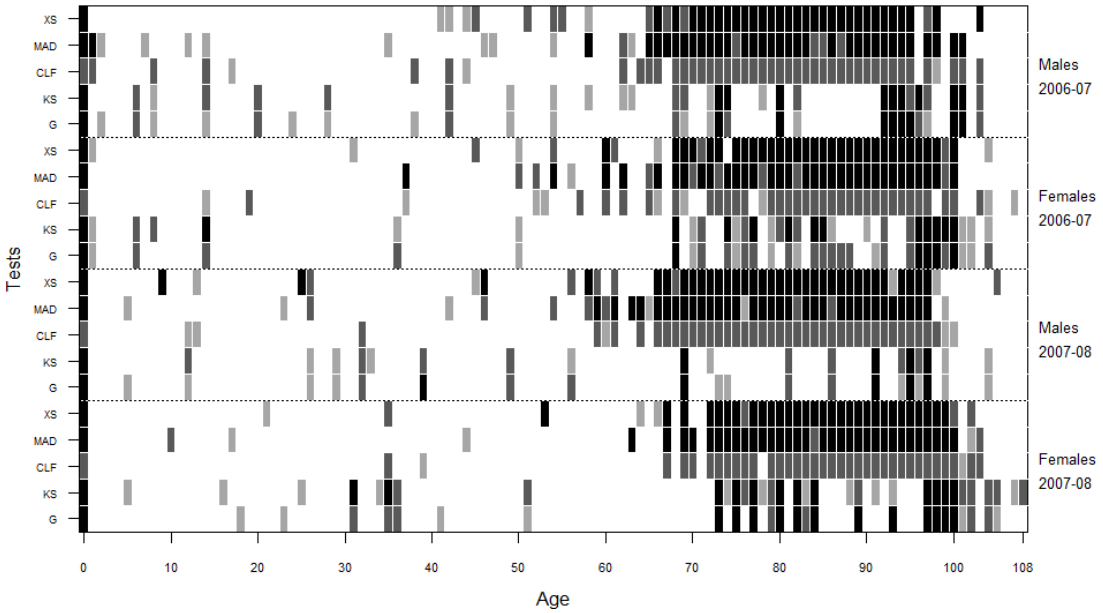

Figure 1A. Death uniform hypothesis tests by gender and age in Spanish population for cohorts dying with age x during years 2006-07 (Lexis quadrilaterals). Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. MAD denotes the Maximum Absolute Deviation test, CLF the Cressie-Loosmore-Ford test, XS the spatial Chi-squared goodness-of-fit test, KS the Kolgomorov-Smirnov test and G the Geometric test. The three first tests are spatial tests and check complete spatial randomness as null hypothesis after a representation of events in the Lexis space. KS and G tests are nonparametric tests and check whether sample exposed-to-risk times are compatible with the assumed probability distributions.

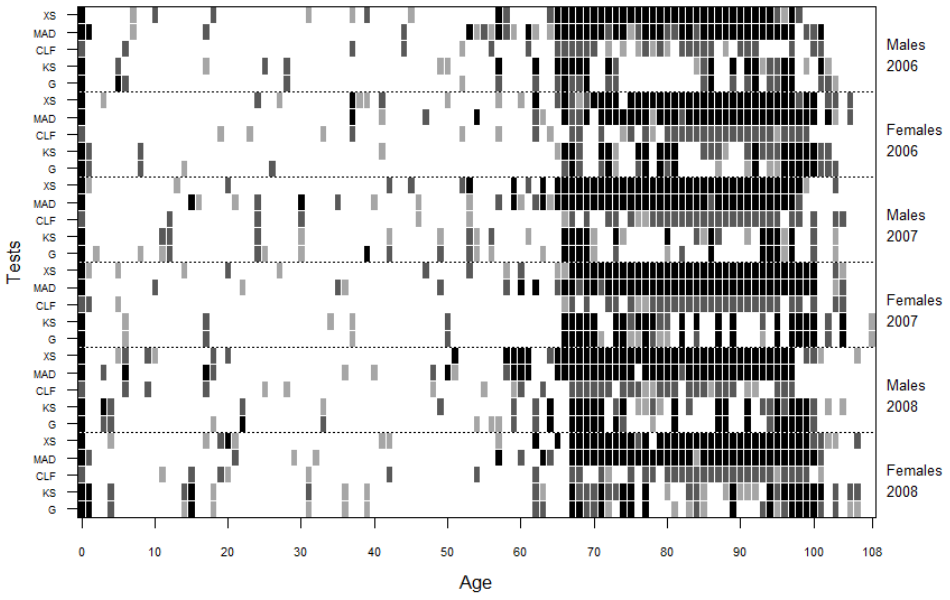

Figure 2A. Death uniform hypothesis tests by gender and age in Spanish population for periods 2006 to 2008 (Lexis cells). Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. MAD denotes the Maximum Absolute Deviation test, CLF the Cressie-Loosmore-Ford test, XS the spatial Chi-squared goodness-of-fit test, KS the Kolgomorov-Smirnov test and G the Geometric test. The three first tests are spatial tests and check complete spatial randomness as null hypothesis after a representation of events in the Lexis space. KS and G tests are

nonparametric tests and check whether sample exposed-to-risk times are compatible with the assumed probability distributions.

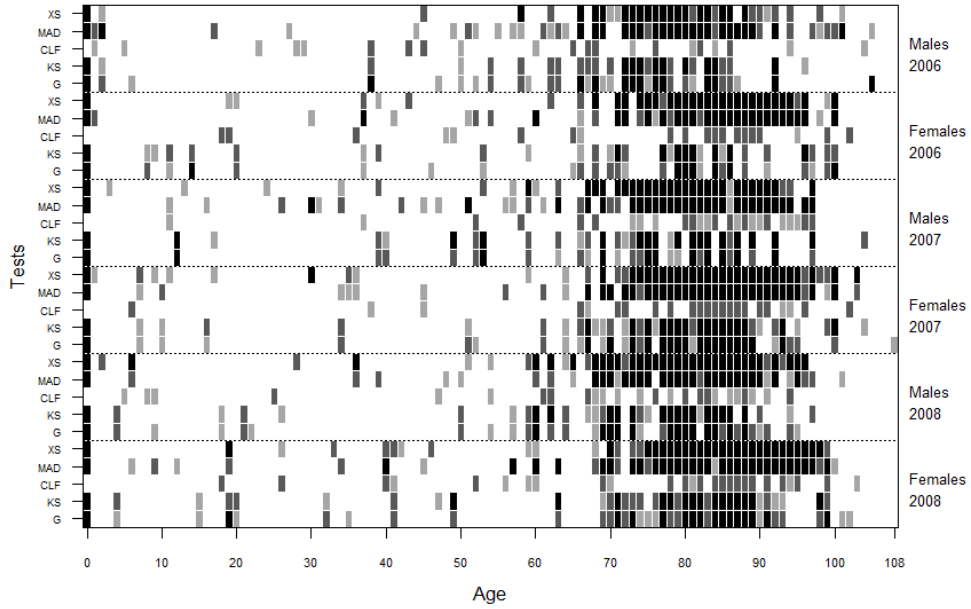

Figure 3A. Death uniform hypothesis tests by gender and age in Spanish population for people dying in Lexis lower triangles in years 2006 to 2008. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. MAD denotes the Maximum Absolute Deviation test, CLF the Cressie-Loosmore-Ford test, XS the spatial Chi-squared goodness-of-fit test, KS the Kolgomorov-Smirnov test and G the Geometric test. The three first tests are spatial tests and check complete spatial randomness as null hypothesis after a representation of events in the Lexis space. KS and G tests are nonparametric tests and check whether sample exposed-to-risk times are compatible with the assumed probability distributions.

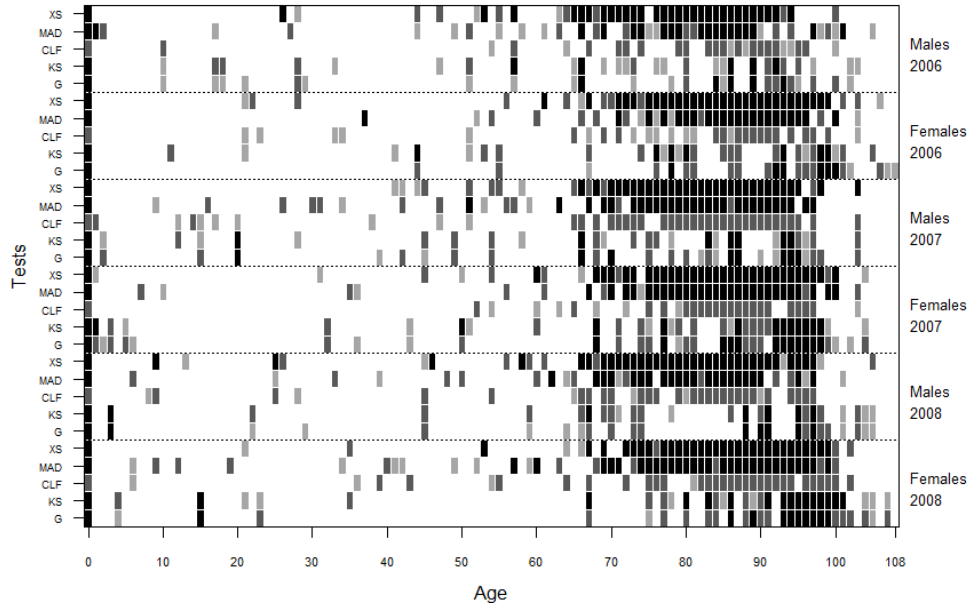

Figure 4A. Death uniform hypothesis tests by gender and age in Spanish population for people dying in Lexis upper triangles during years 2006 to 2008. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. MAD denotes the Maximum Absolute Deviation test, CLF the Cressie-Loosmore-Ford test, XS the spatial Chi-squared goodness-of-fit test, KS the Kolgomorov-Smirnov test and G the Geometric test. The three first tests are spatial tests and

check complete spatial randomness as null hypothesis after a representation of events in the Lexis space. KS and G tests are nonparametric tests and check whether sample exposed-to-risk times are compatible with the assumed probability distributions.

EMIGRANT UNIFORM HYPOTHESIS TESTS

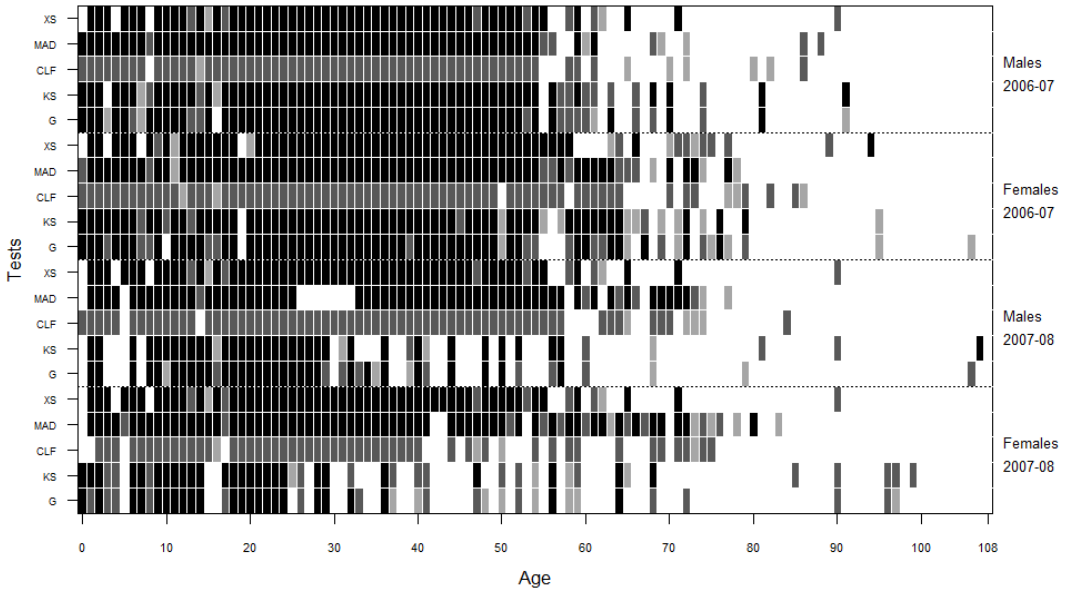

Figure 5A. Emigrant uniform hypothesis tests by gender and age in Spanish population for cohorts emigrating with age x in years 2006-07 and 2007-08 (Lexis quadrilaterals). Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. MAD denotes the Maximum Absolute Deviation test, CLF the Cressie-Loosmore-Ford test, XS the spatial Chi-squared goodness-of-fit test, KS the Kolgomorov-Smirnov test and G the Geometric test. The three first tests are spatial tests and check complete spatial randomness as null hypothesis after a representation of events in the Lexis space. KS and G tests are nonparametric tests and check whether sample exposed-to-risk times are compatible with the assumed probability distributions.

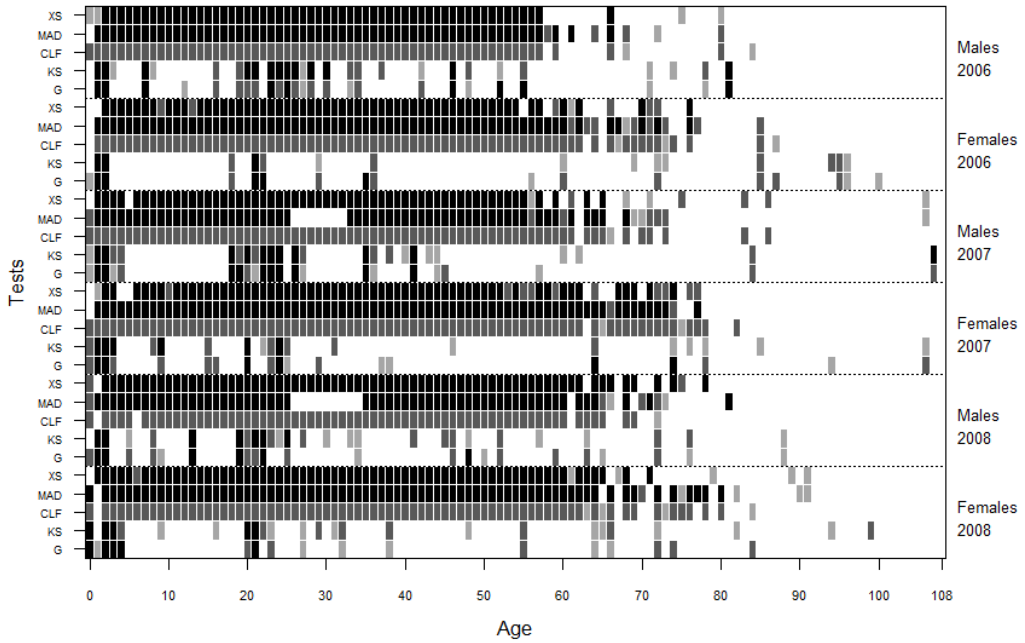

Figure 6A. Uniform hypothesis tests for emigrant events by gender and age in Spanish population during periods from 2006 to 2008 (Lexis cells). Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. MAD denotes the Maximum Absolute Deviation test, CLF the Cressie-Loosmore-Ford test, XS the spatial Chi-squared goodness-of-fit test, KS the Kolgomorov-Smirnov test and G the Geometric test. The three first tests are spatial tests and check complete spatial randomness as null hypothesis after a representation of events in the Lexis space. KS and

G tests are nonparametric tests and check whether sample exposed-to-risk times are compatible with the assumed probability distributions.

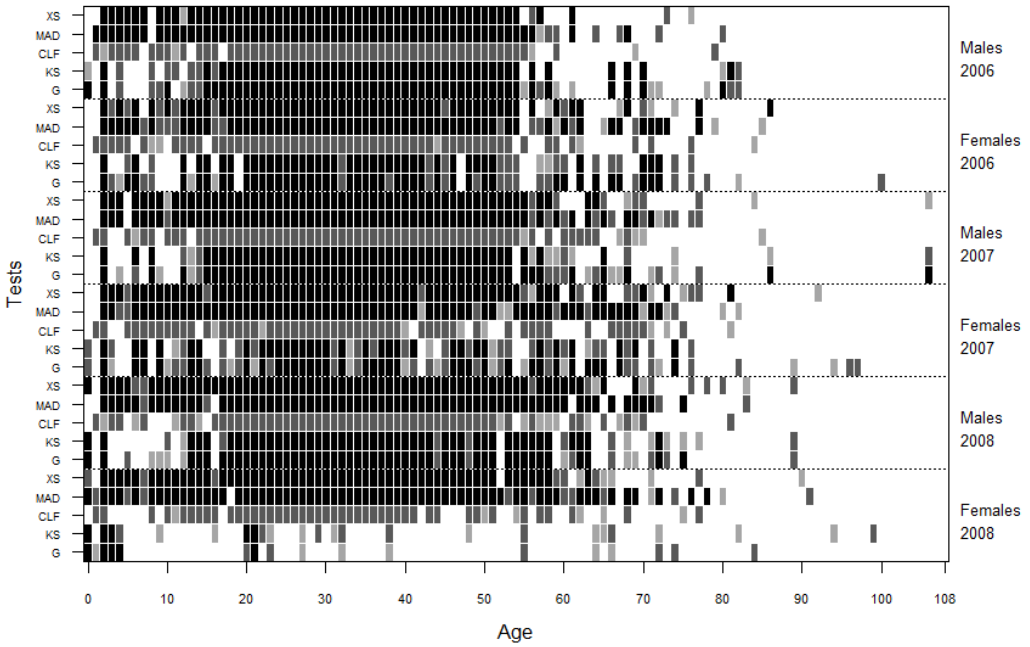

Figure 7A. Uniform hypothesis tests for emigrant events by gender and age in Spanish in Lexis lower triangles in years 2006 to 2008. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. MAD denotes the Maximum Absolute Deviation test, CLF the Cressie-Loosmore-Ford test, XS the spatial Chi-squared goodness-of-fit test, KS the Kolgomorov-Smirnov test and G the Geometric test. The three first tests are spatial tests and check complete spatial randomness as null hypothesis after a representation of events in the Lexis space. KS and G tests are nonparametric tests and check whether sample exposed-to-risk times are compatible with the assumed probability distributions.

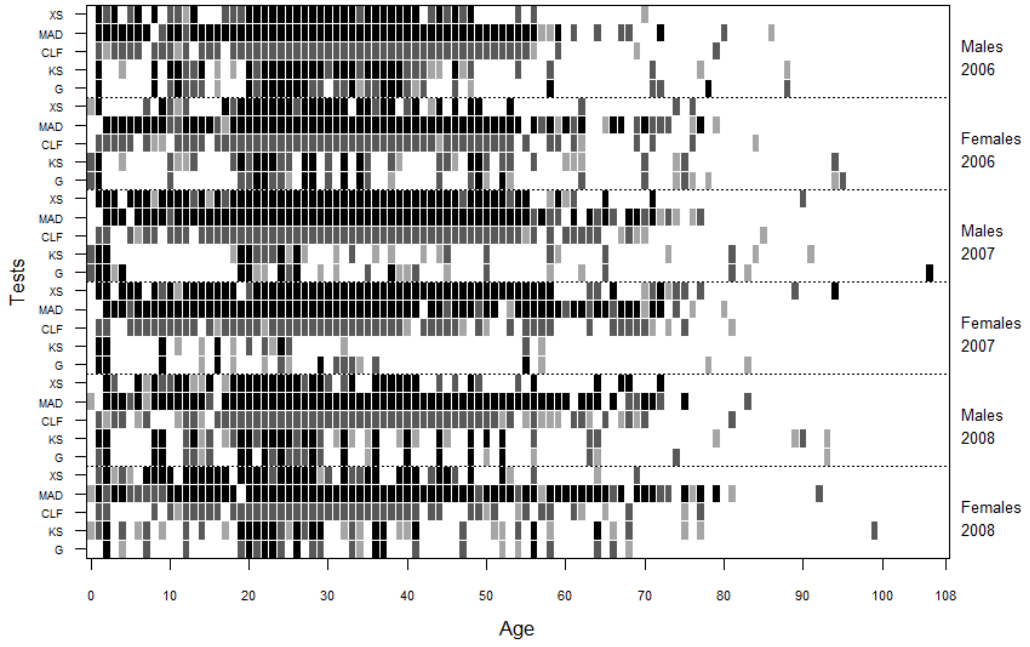

Figure 8A. Uniform hypothesis tests for emigrant events by gender and age in Spanish in Lexis upper triangles in years 2006 to 2008. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. MAD denotes the Maximum Absolute Deviation test,

CLF the Cressie-Loosmore-Ford test, XS the spatial Chi-squared goodness-of-fit test, KS the Kolgomorov-Smirnov test and G the Geometric test. The three first tests are spatial tests and check complete spatial randomness as null hypothesis after a representation of events in the Lexis space. KS and G tests are nonparametric tests and check whether sample exposed-to-risk times are compatible with the assumed probability distributions.

IMMIGRANT UNIFORM HYPOTHESIS TESTS

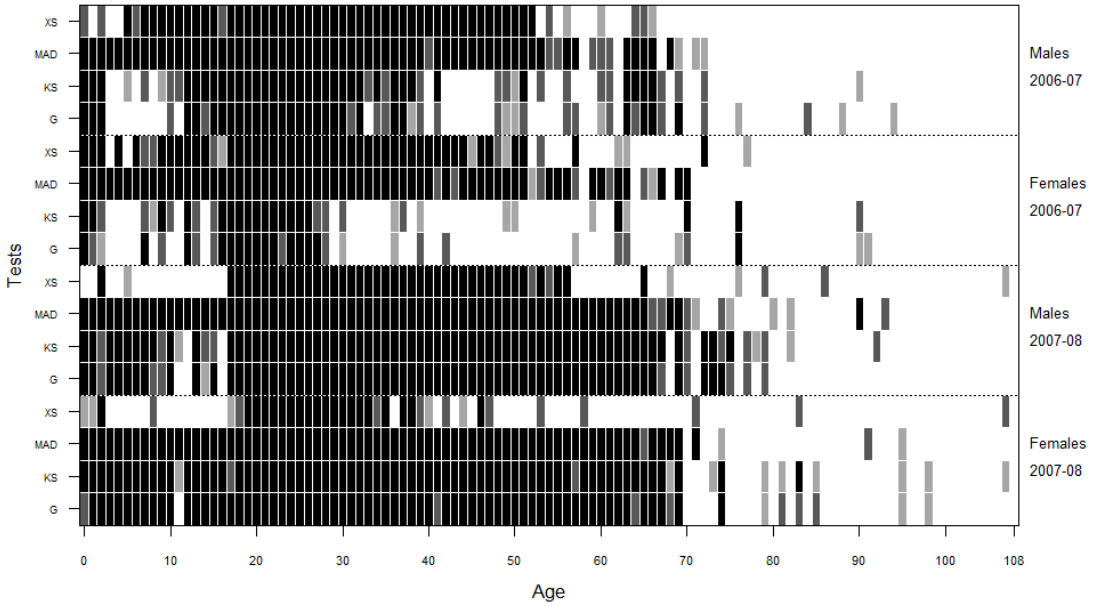

Figure 9A. Immigrant uniform hypothesis tests by gender and age in Spanish population for cohorts immigrating with age x in years 2006-07 and 2007-08 (Lexis quadrilaterals). Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. MAD denotes the Maximum Absolute Deviation test, XS the spatial Chi-squared goodness-of-fit test, KS the Kolgomorov-Smirnov test and G the Geometric test. CLF tests were not performed in this case because the R function dclf.test was unable to handle the large number of immigrant events occurring in the majority of surfaces. The two first tests are spatial tests and check complete spatial randomness as null hypothesis after a representation of events in the Lexis space. KS and G tests are nonparametric tests and check whether sample exposed-to-risk times are compatible with the assumed probability distributions.

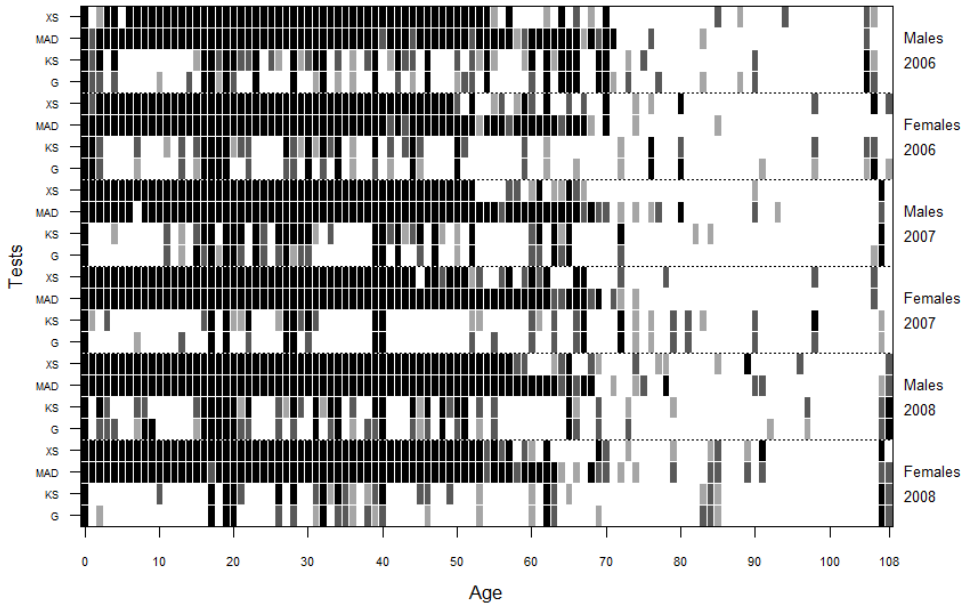

Figure 10A. Uniform hypothesis tests for immigrant events by gender and age in Spanish population during periods 2006 to 2008 (Lexis cells). Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. MAD denotes the Maximum Absolute Deviation test, XS the spatial Chi-squared goodness-of-fit test, KS the Kolgomorov-Smirnov test and G the Geometric test. CLF tests were not performed in this case because the R function dclf.test was unable to

handle the large number of immigrant events occurring in the majority of surfaces. The two first tests are spatial tests and check complete spatial randomness as null hypothesis after a representation of events in the Lexis space. KS and G tests are nonparametric tests and check whether sample exposed-to-risk times are compatible with the assumed probability distributions.

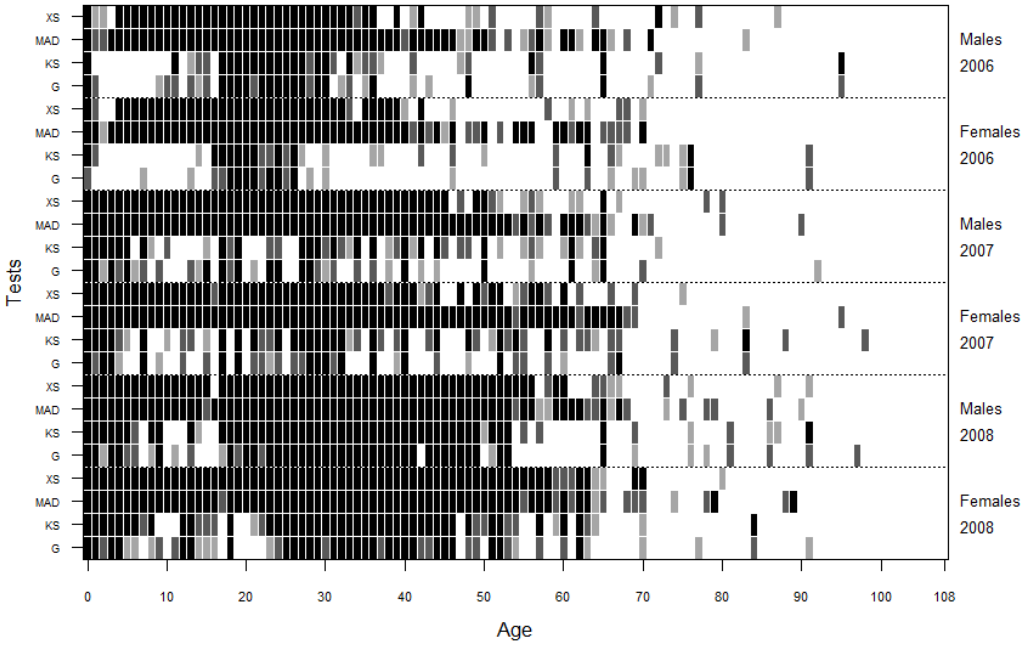

Figure 11A. Uniform hypothesis tests for immigrant events by gender and age in Spanish in Lexis lower triangles in years 2006 to 2008. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. MAD denotes the Maximum Absolute Deviation test, XS the spatial Chi-squared goodness-of-fit test, KS the Kolgomorov-Smirnov test and G the Geometric test. CLF tests were not performed in this case because the R function dclf.test was unable to handle the large number of immigrant events occurring in the majority of surfaces. The two first tests are spatial tests and check complete spatial randomness as null hypothesis after a representation of events in the Lexis space. KS and G tests are nonparametric tests and check whether sample exposed-to-risk times are compatible with the assumed probability distributions.

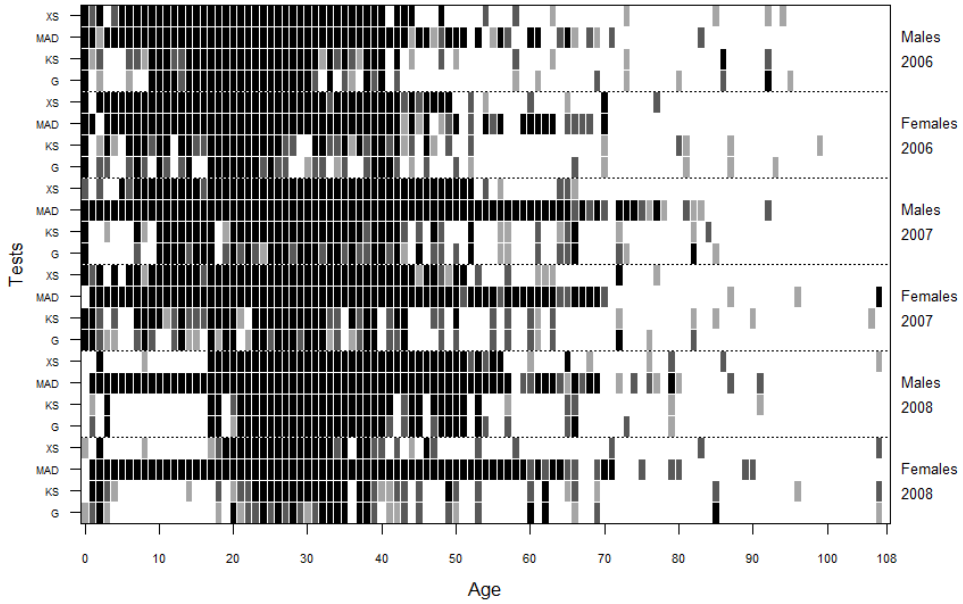

Figure 12A. Uniform hypothesis tests for immigrant events by gender and age in Spanish in Lexis upper triangles in years 2006 to 2008. Black, dark grey and light grey colours indicate rejection of null hypotheses

at, respectively, 0.01, 0.05 and 0.1 significant levels. MAD denotes the Maximum Absolute Deviation test, XS the spatial Chi-squared goodness-of-fit test, KS the Kolgomorov-Smirnov test and G the Geometric test. CLF tests were not performed in this case because the R function dclf.test was unable to handle the large number of immigrant events occurring in the majority of surfaces. The two first tests are spatial tests and check complete spatial randomness as null hypothesis after a representation of events in the Lexis space. KS and G tests are nonparametric tests and check whether sample exposed-to-risk times are compatible with the assumed probability distributions.

HYPOTHESIS TESTS. ADDITIONAL FIGURES

AVERAGE TIME EXPOSED TO RISK AND NUMBER OF EVENTS: DEATH MALES.

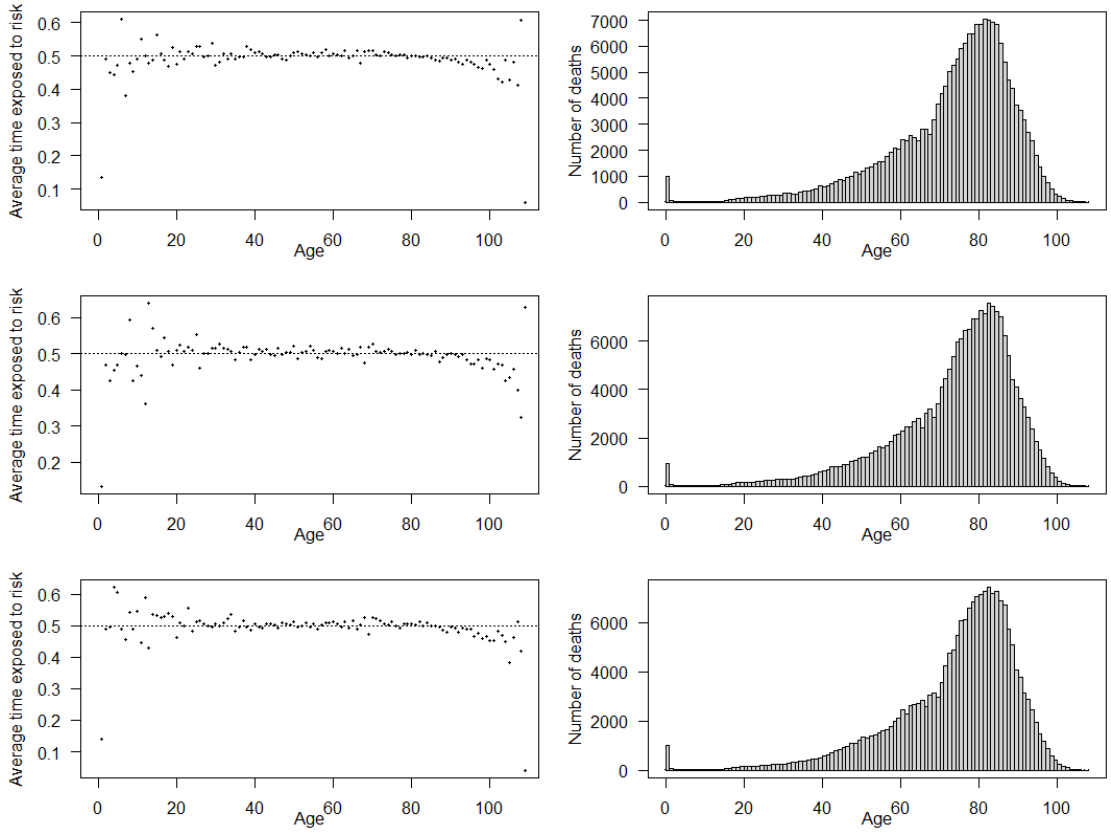

Figure 13A. Average number of years lived at age of dying and number of deaths by age for males dying during years 2006 (upper), 2007 (middle) and 2008 (lower).

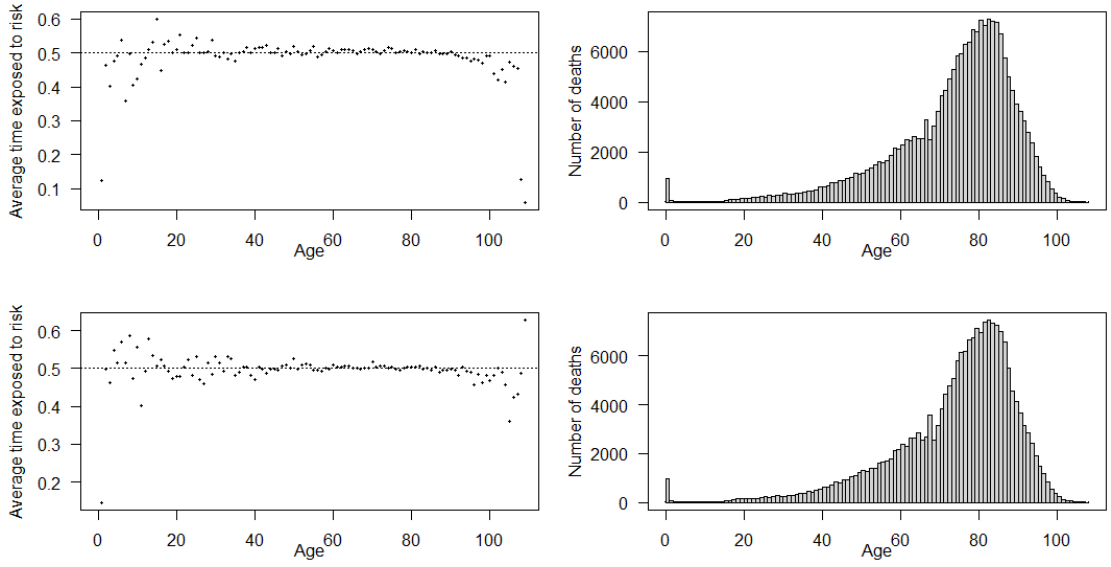

Figure 14A. Average number of years lived at age of dying and number of deaths by age for cohort of males dying with age x in years 2006-07 (upper) and 2007-08 (lower).

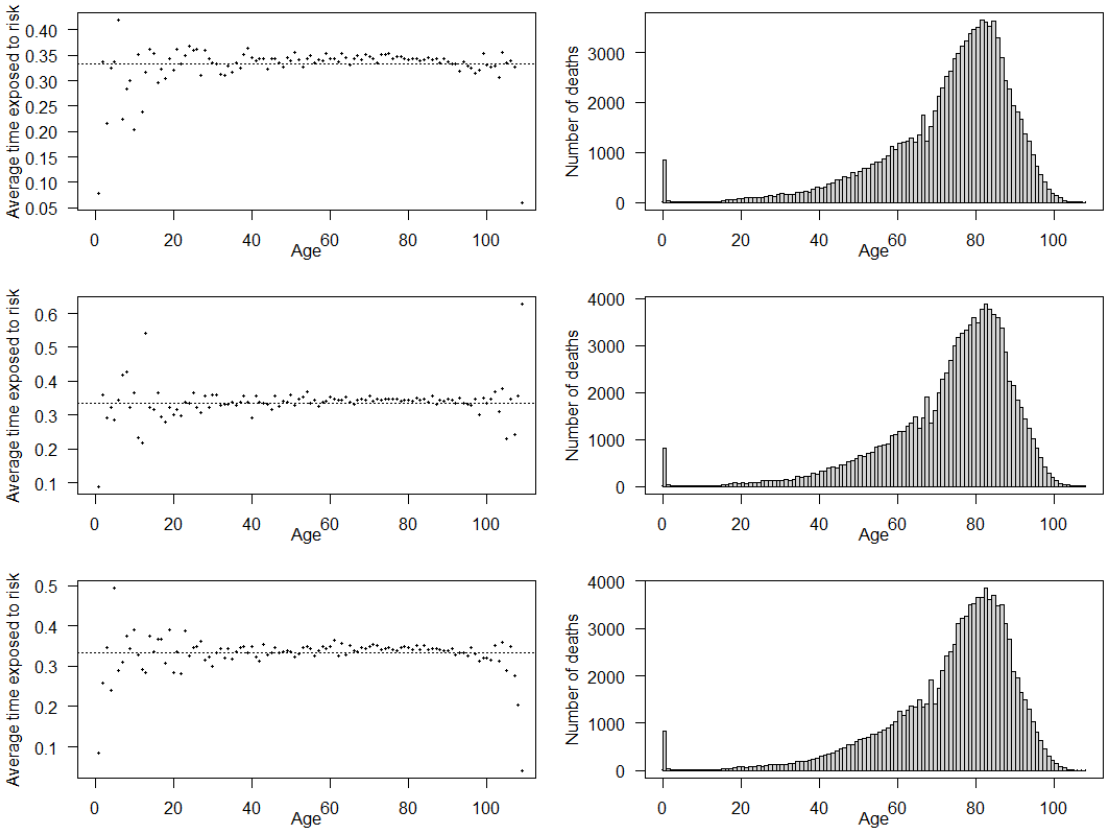

Figure 15A. Average number of years lived at age of dying and number of deaths by age for males dying in lower triangles during years 2006 (upper), 2007 (middle) and 2008 (lower).

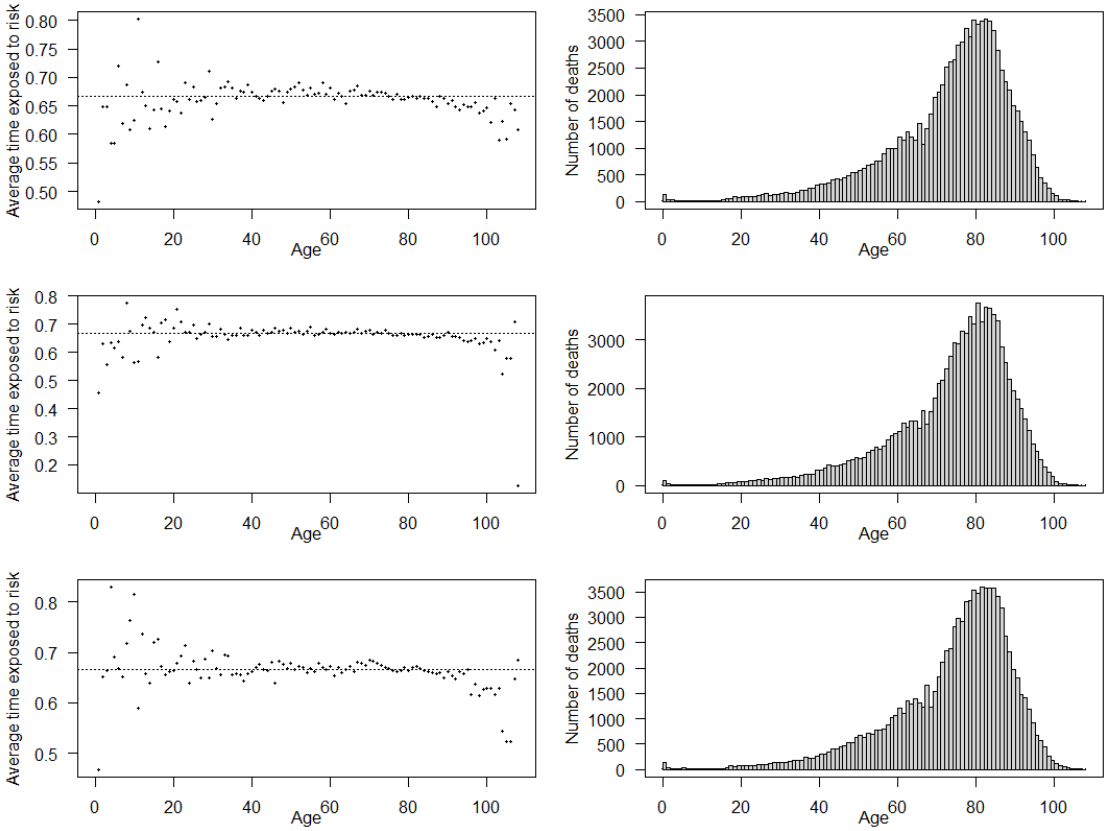

Figure 16A. Average number of years lived at age of dying and number of deaths by age for males dying in upper triangles during years 2006 (upper), 2007 (middle) and 2008 (lower).

AVERAGE TIME EXPOSED TO RISK AND NUMBER OF EVENTS: DEATH FEMALES.

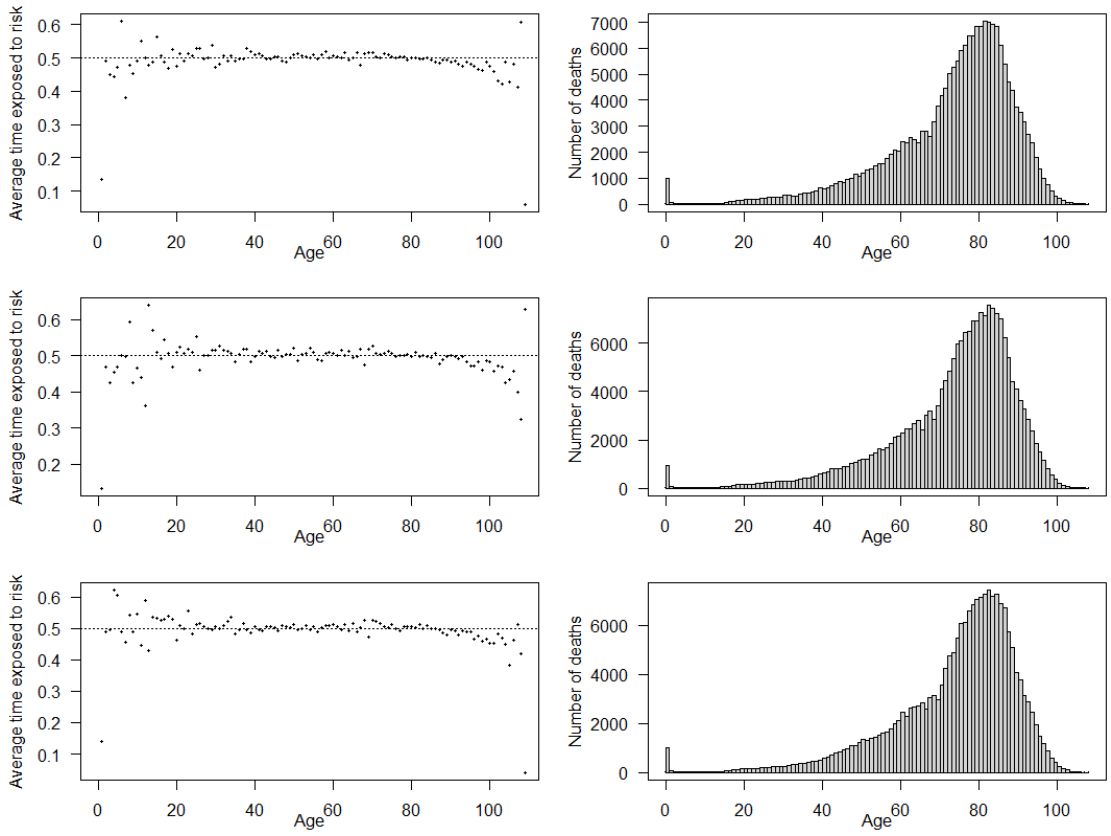

Figure 17A. Average number of years lived at age of dying and number of deaths by age for females dying during years 2006 (upper), 2007 (middle) and 2008 (lower).

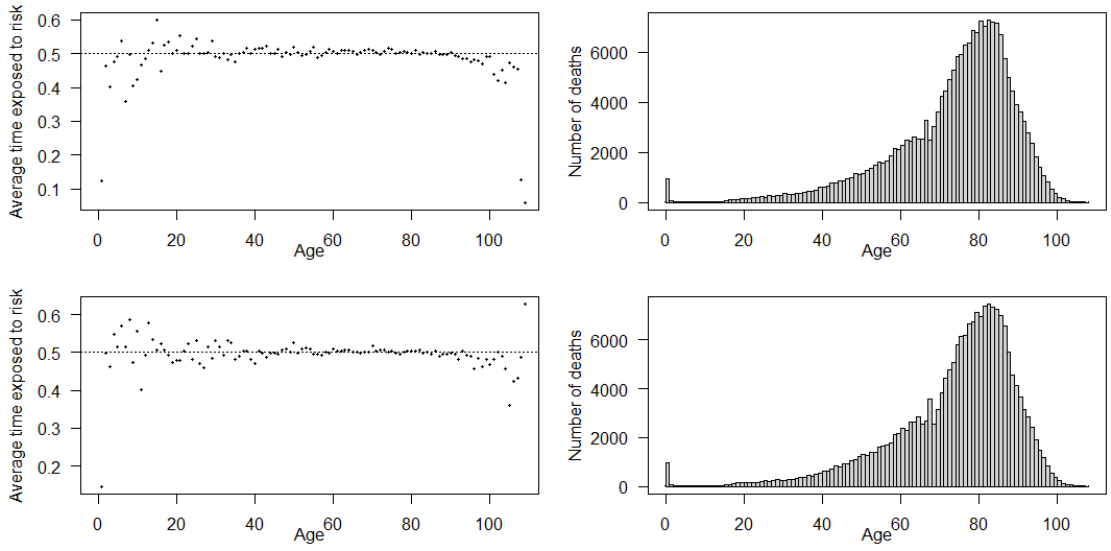

Figure 18A. Average number of years lived at age of dying and number of deaths by age for cohort of females dying with age x in years 2006-07 (upper) and 2007-08 (lower).

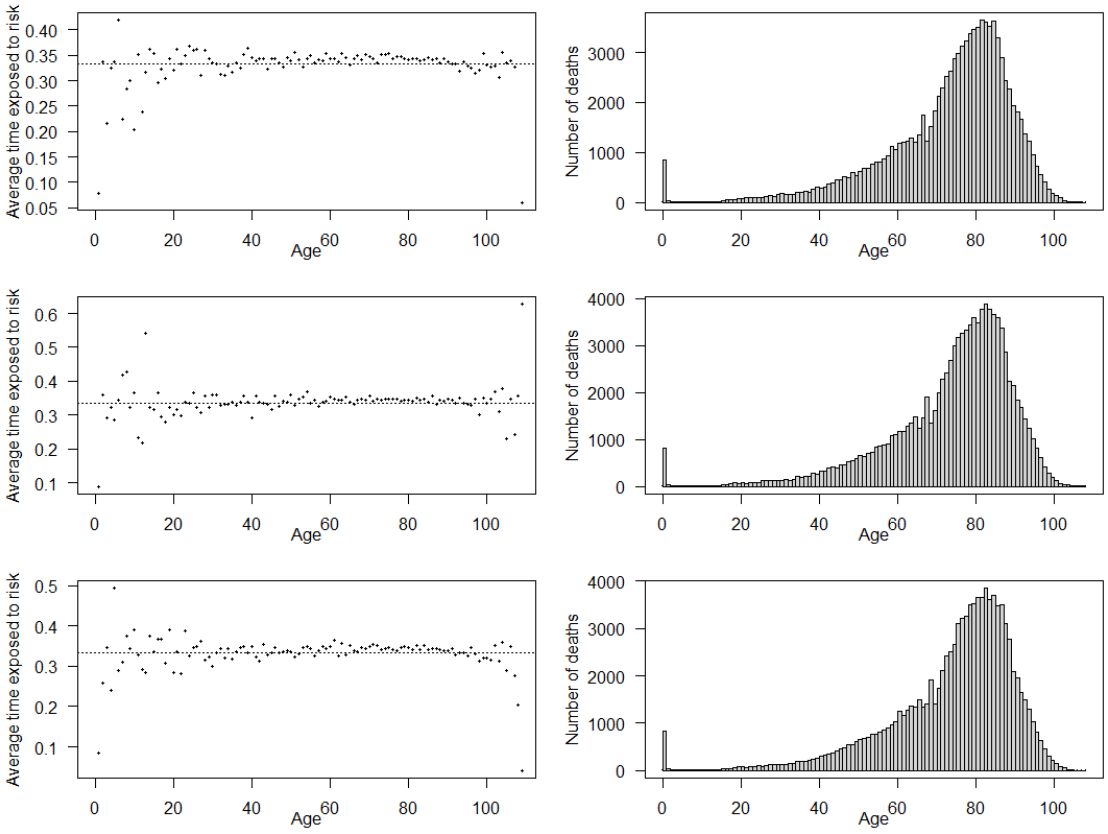

Figure 19A. Average number of years lived at age of dying and number of deaths by age for males dying in lower triangles during years 2006 (upper), 2007 (middle) and 2008 (lower).

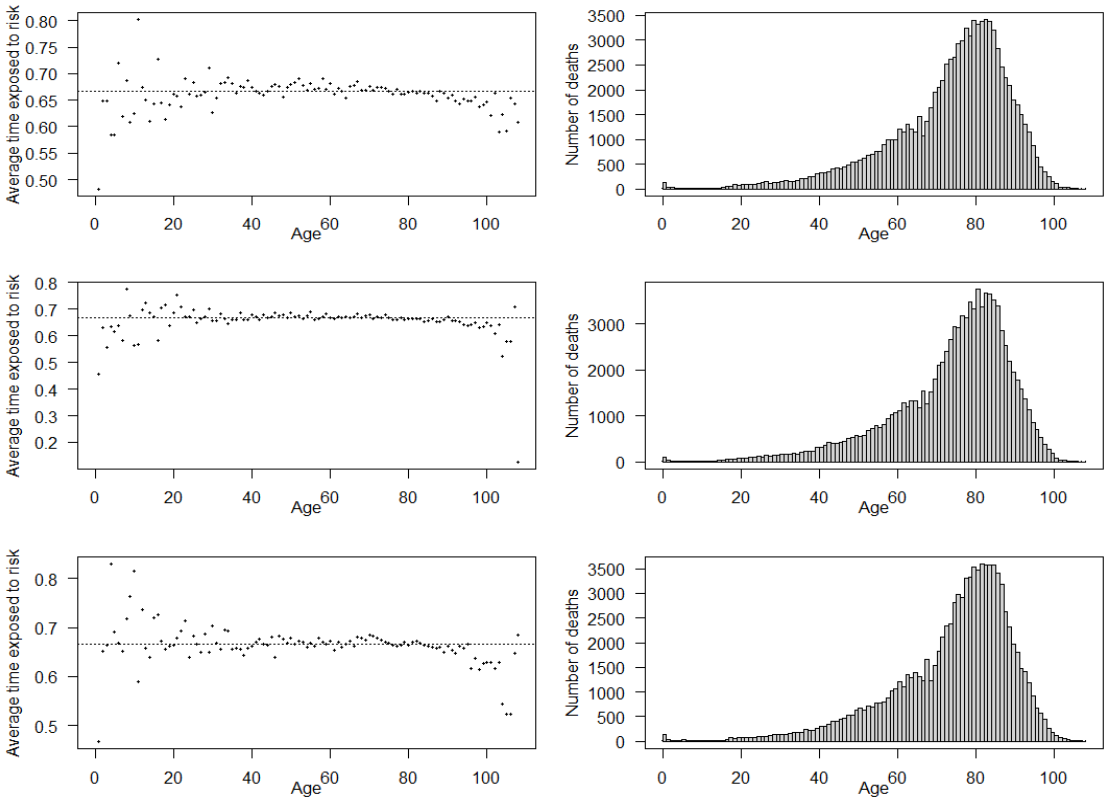

Figure 20A. Average number of years lived at age of dying and number of deaths by age for females dying in upper triangles during years 2006 (upper), 2007 (middle) and 2008 (lower).

AVERAGE TIME EXPOSED TO RISK AND NUMBER OF EVENTS: EMIGRANT MALES.

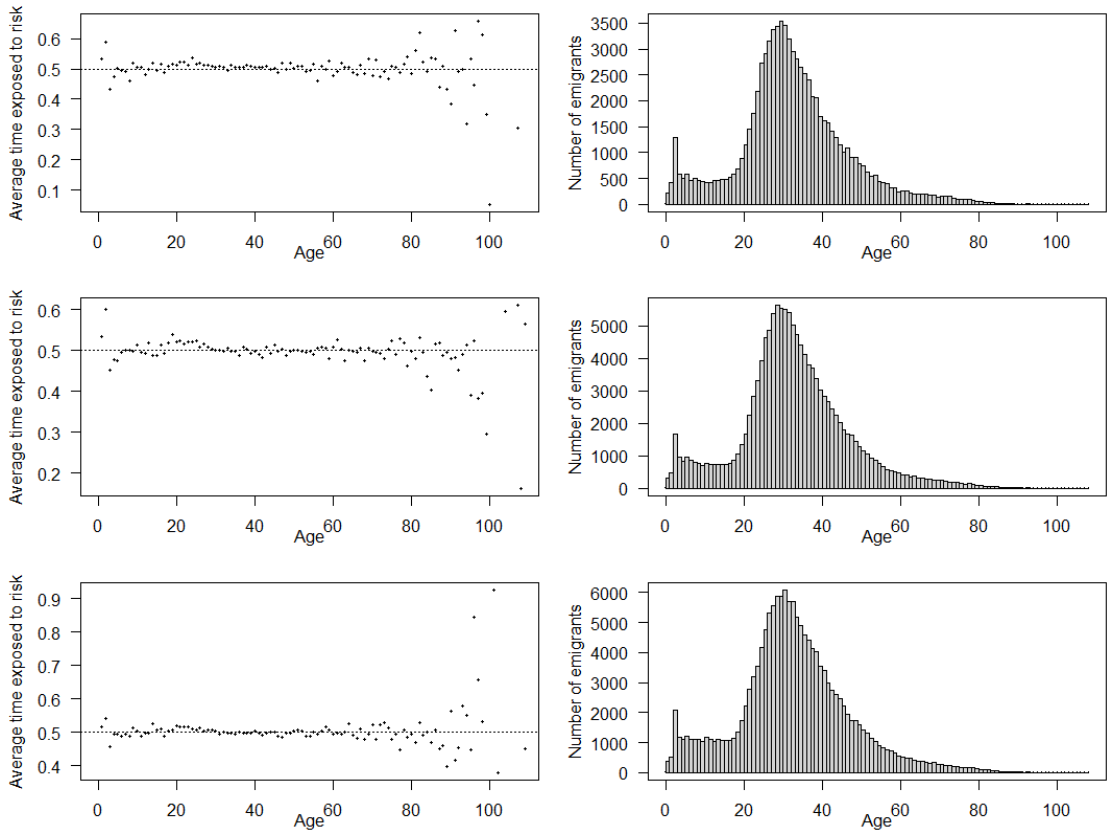

Figure 21A. Average number of years exposed to risk and number of emigrants by age for males emigrating during years 2006 (upper), 2007 (middle) and 2008 (lower).

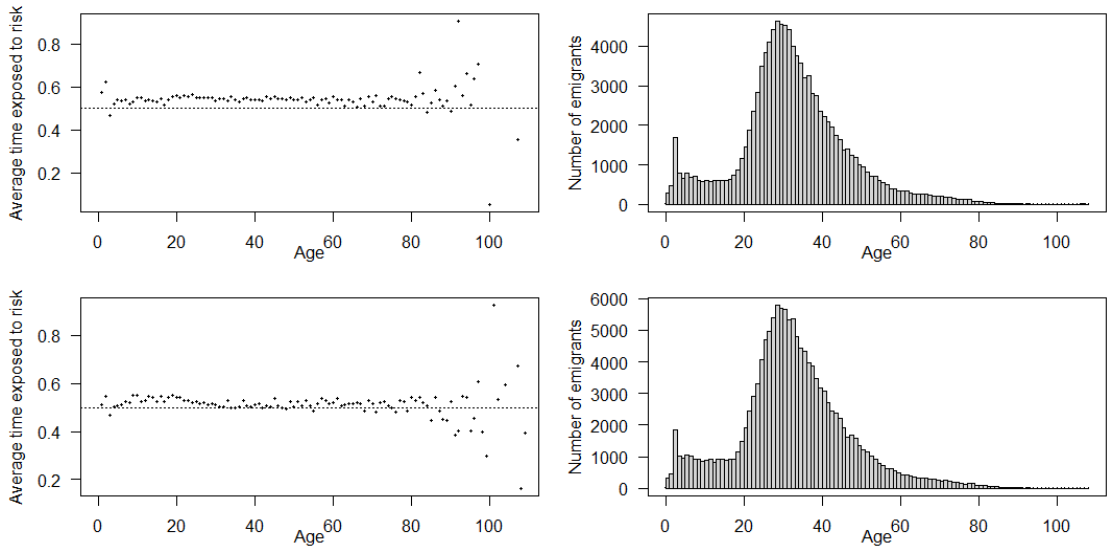

Figure 22A. Average number of years exposed to risk and number of emigrants by age for cohort of males emigrating with age x in years 2006-07 (upper) and 2007-08 (lower).

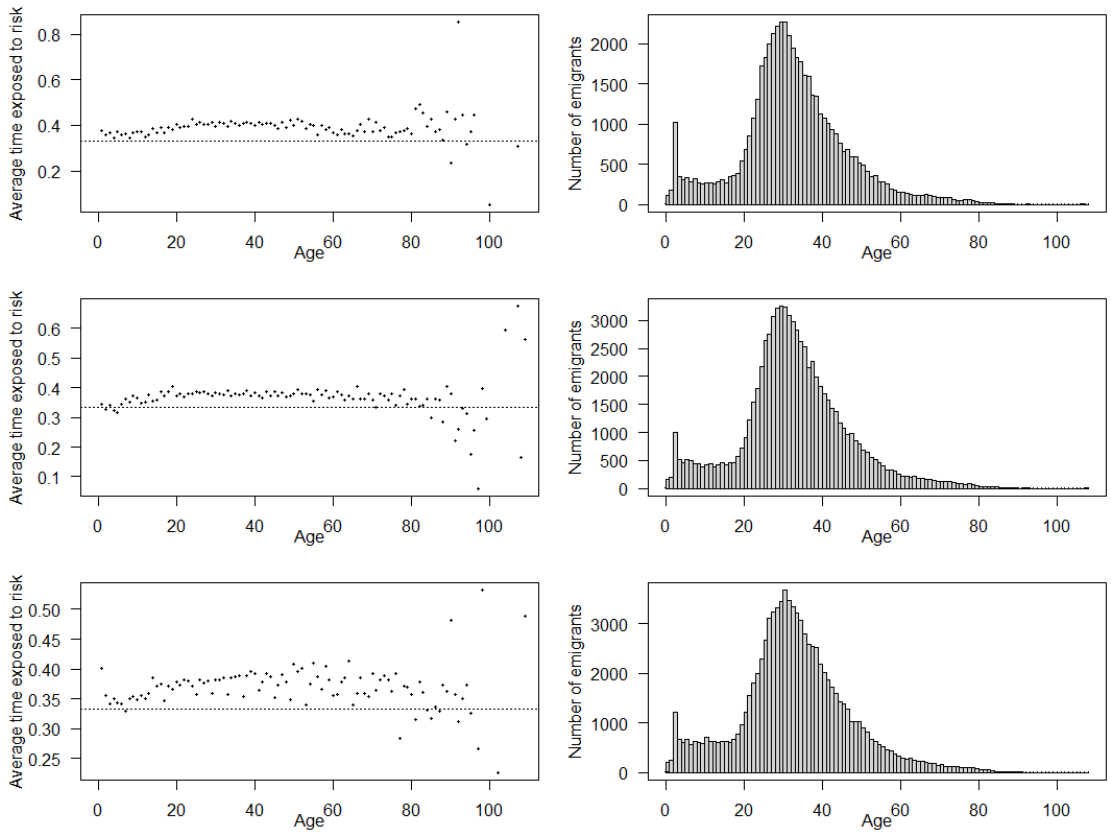

Figure 23A. Average number of years exposed to risk and number of emigrants by age for males emigrating in lower triangles during years 2006 (upper), 2007 (middle) and 2008 (lower).

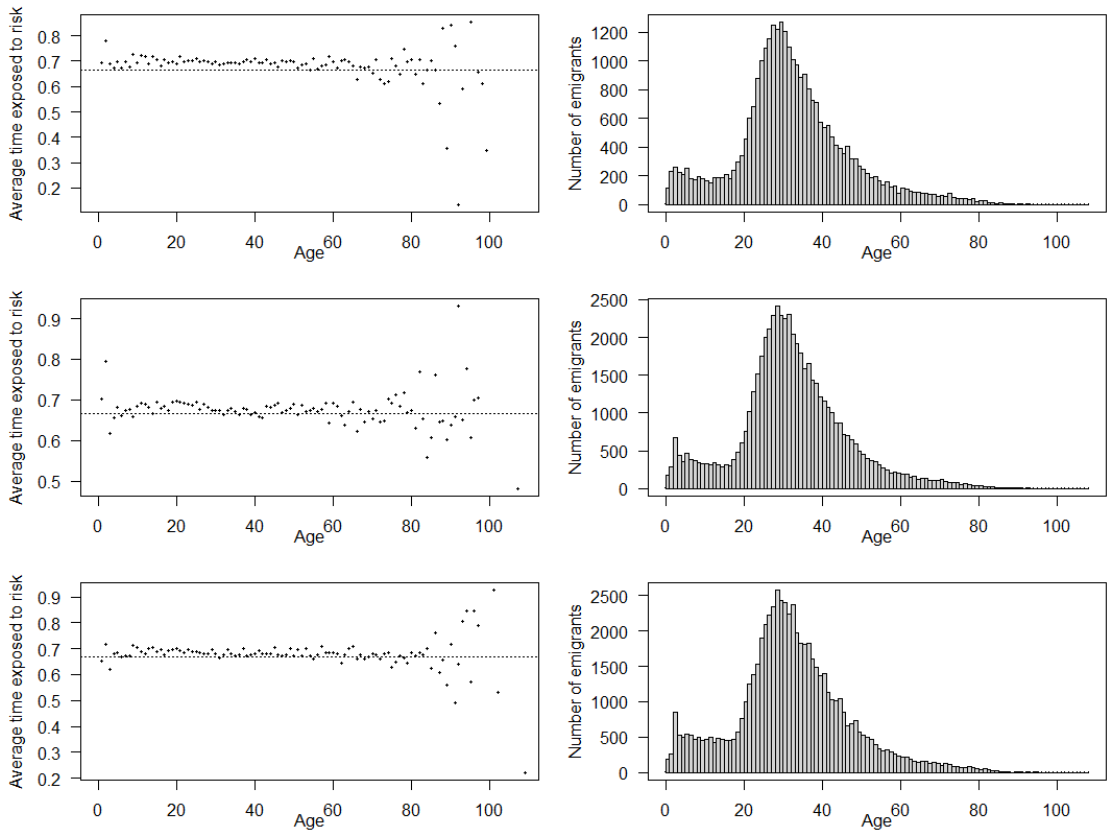

Figure 24A. Average number of years exposed to risk and number of emigrants by age for males emigrating in upper triangles in years 2006 (upper), 2007 (middle) and 2008 (lower).

AVERAGE TIME EXPOSED TO RISK AND NUMBER OF EVENTS: EMIGRANT FEMALES.

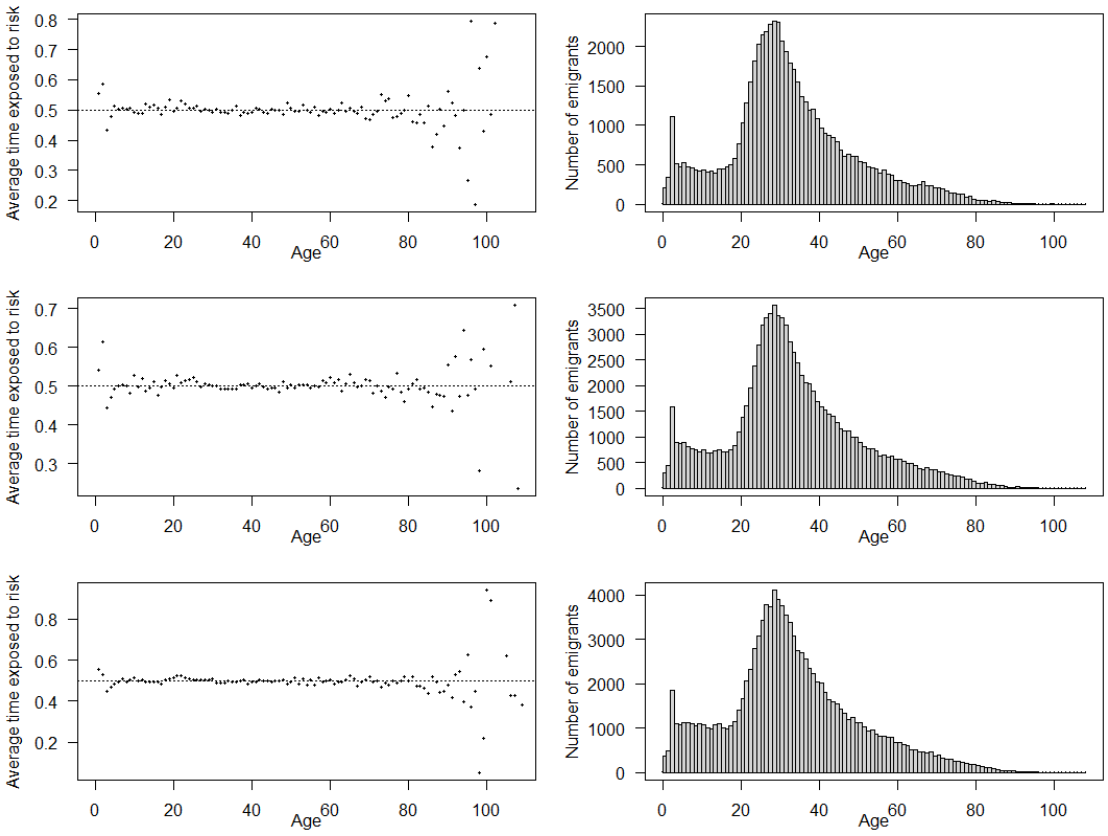

Figure 25A. Average number of years exposed to risk and number of emigrants by age for females emigrating during years 2006 (upper), 2007 (middle) and 2008 (lower).

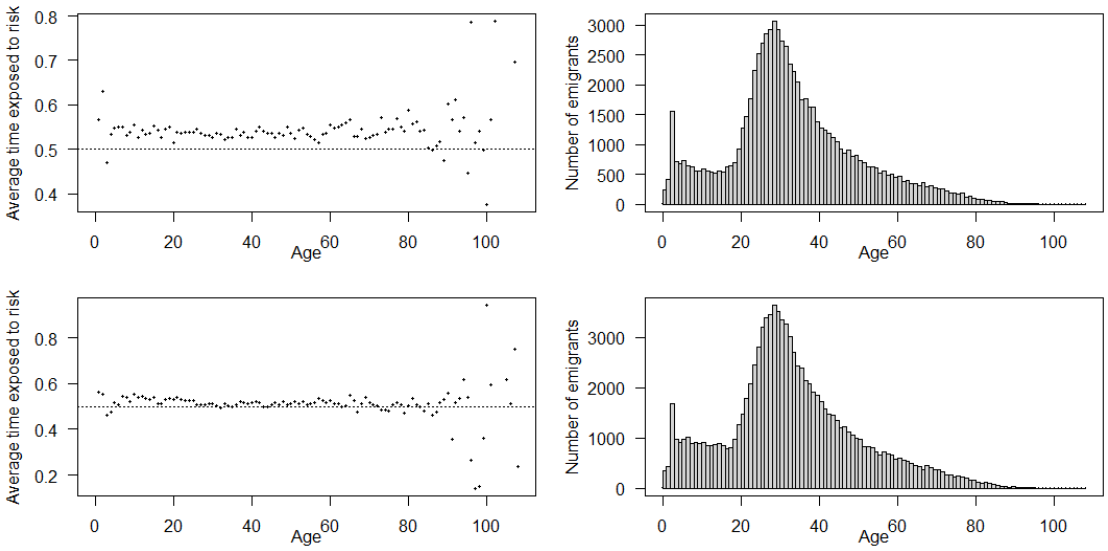

Figure 26A. Average number of years exposed to risk and number of emigrants by age for cohort of females emigrating with age x in years 2006-07 (upper) and 2007-08 (lower).

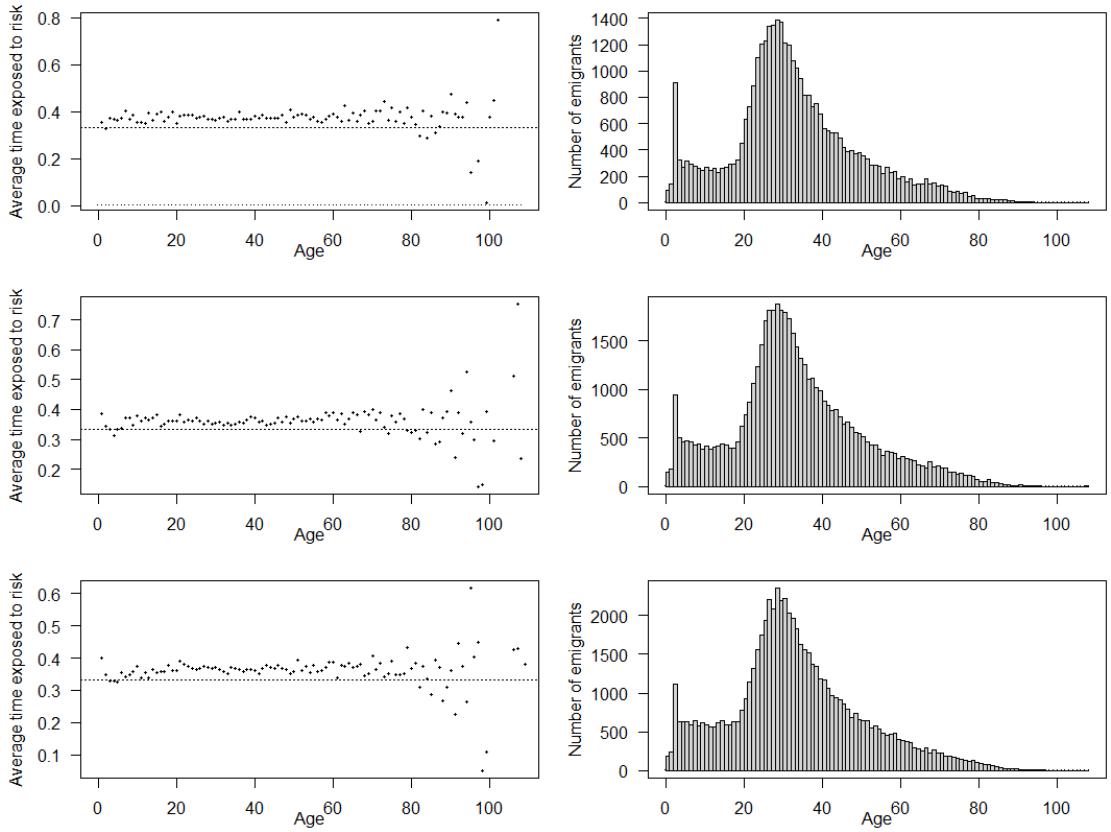

Figure 27A. Average number of years exposed to risk and number of emigrants by age for females emigrating in lower triangles in years 2006 (upper), 2007 (middle) and 2008 (lower).

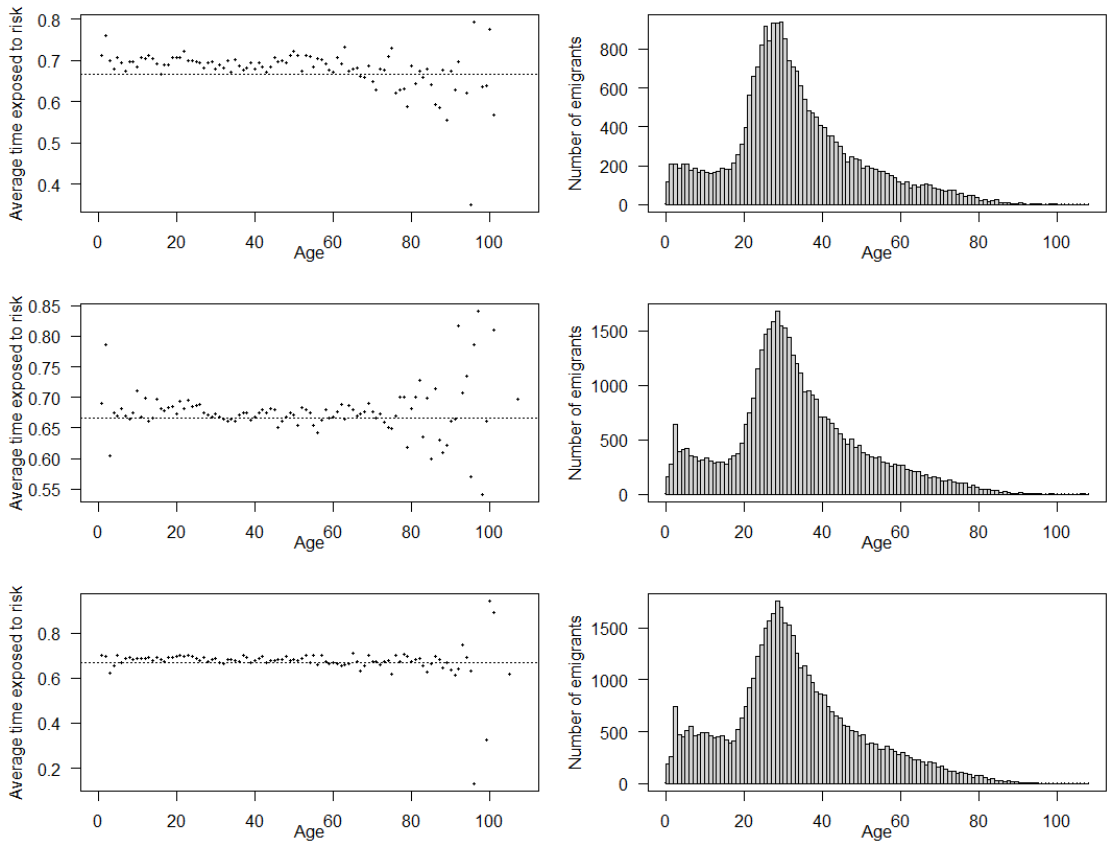

Figure 28A. Average number of years exposed to risk and number of emigrants by age for females emigrating in upper triangles in years 2006 (upper), 2007 (middle) and 2008 (lower).

AVERAGE TIME EXPOSED TO RISK AND NUMBER OF EVENTS: IMMIGRANT MALES.

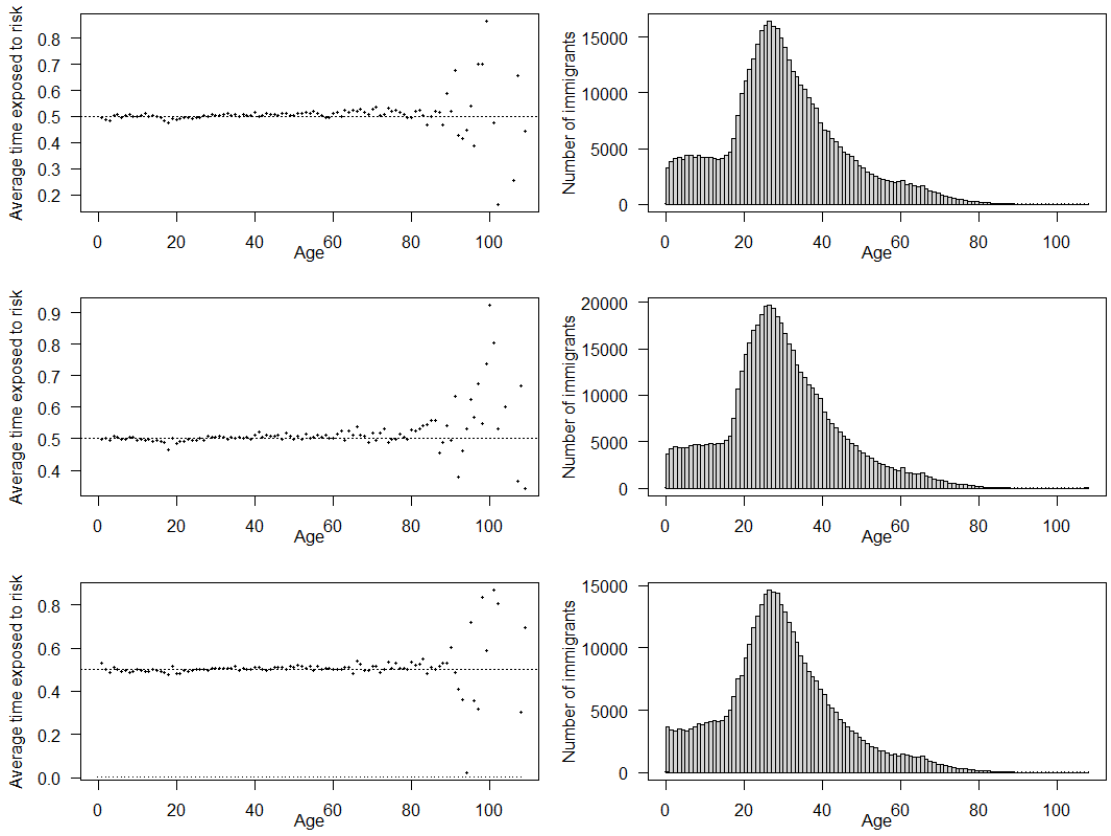

Figure 29A. Average number of years exposed to risk and number of immigrants by age for males immigrating during years 2006 (upper), 2007 (middle) and 2008 (lower).

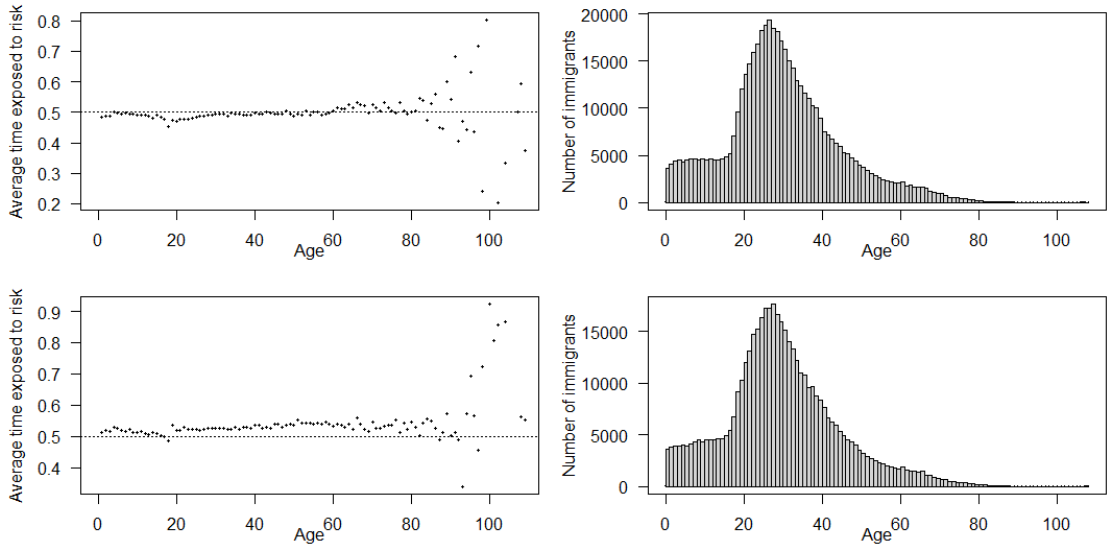

Figure 30A. Average number of years exposed to risk and number of immigrants by age for cohort of males immigrating with age x in years 2006-07 (upper) and 2007-08 (lower).

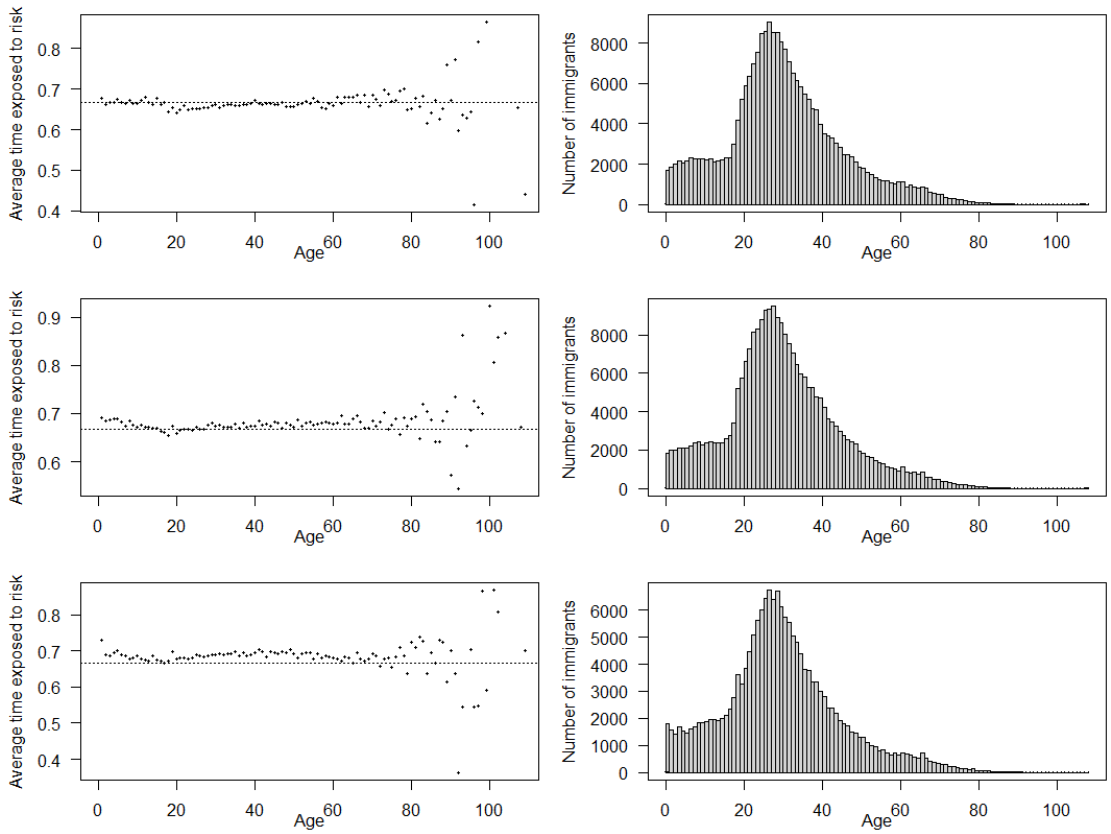

Figure 31A. Average number of years exposed to risk and number of immigrants by age for males immigrating in lower triangles in years 2006 (upper), 2007 (middle) and 2008 (lower).

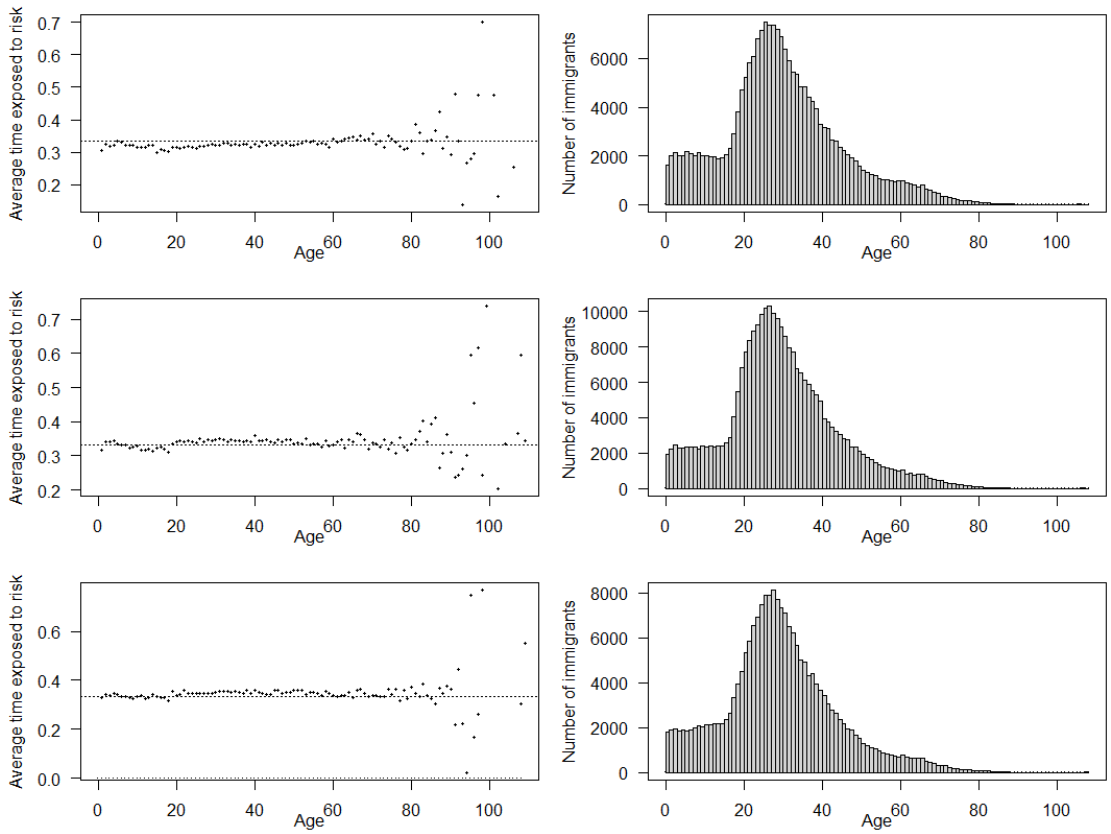

Figure 32A. Average number of years exposed to risk and number of immigrants by age for males immigrating in upper triangles in years 2006 (upper), 2007 (middle) and 2008 (lower).

AVERAGE TIME EXPOSED TO RISK AND NUMBER OF EVENTS: IMMIGRANT FEMALES.

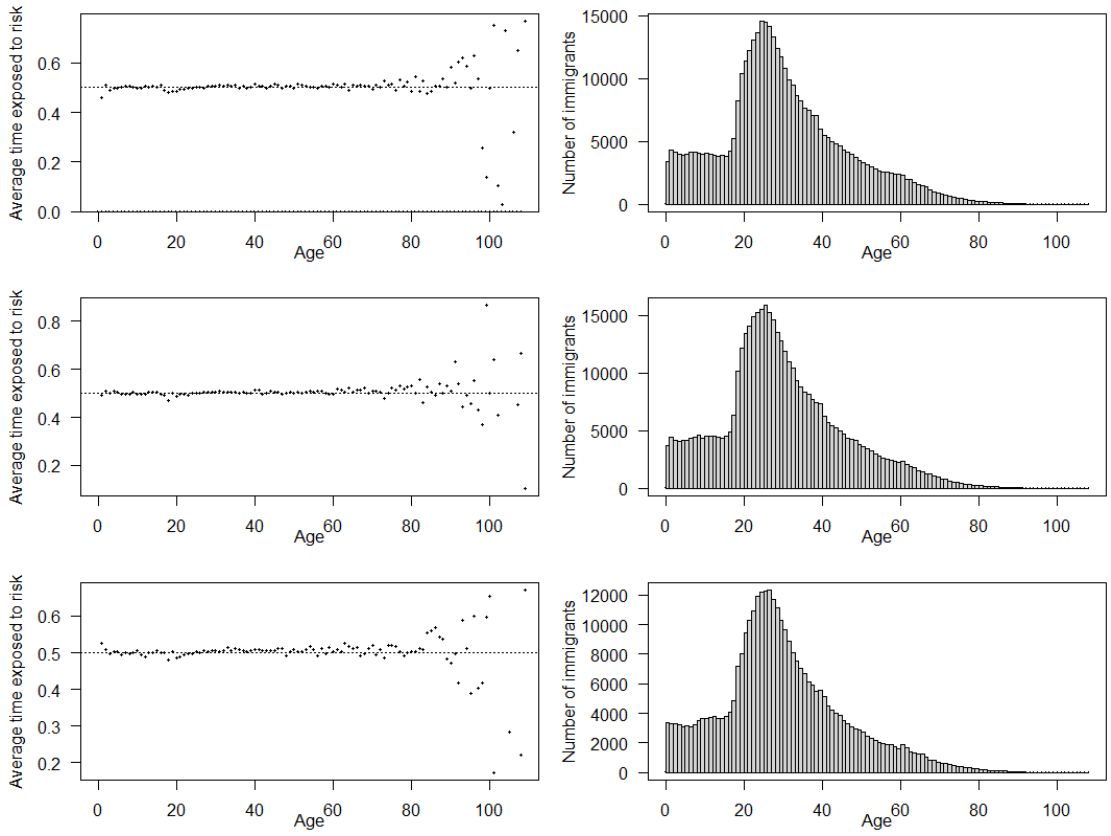

Figure 33A. Average number of years exposed to risk and number of immigrants by age for females immigrating during years 2006 (upper), 2007 (middle) and 2008 (lower).

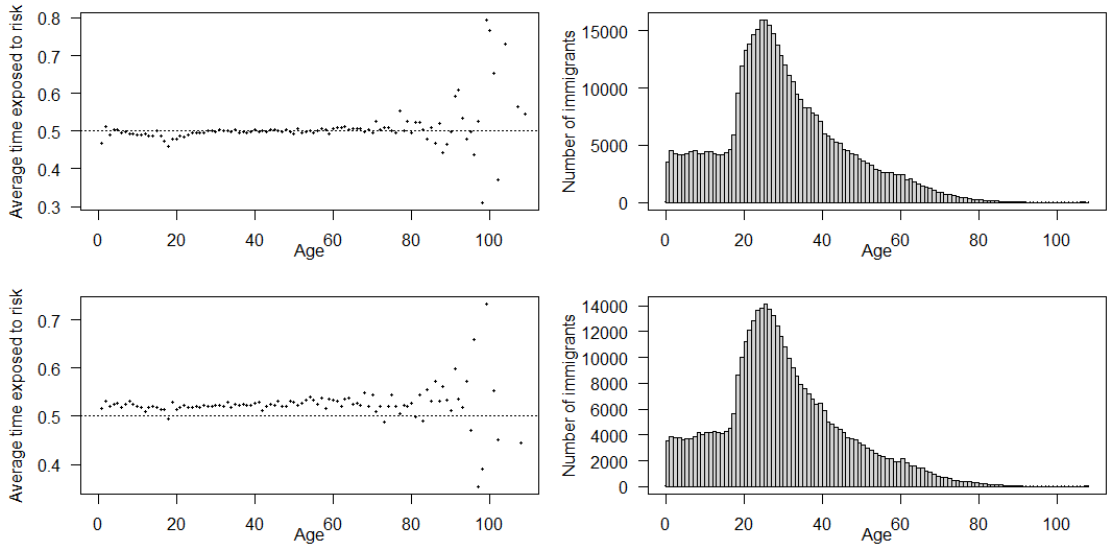

Figure 34A. Average number of years exposed to risk and number of immigrants by age for cohort of females immigrating with age x in years 2006-07 (upper) and 2007-08 (lower).

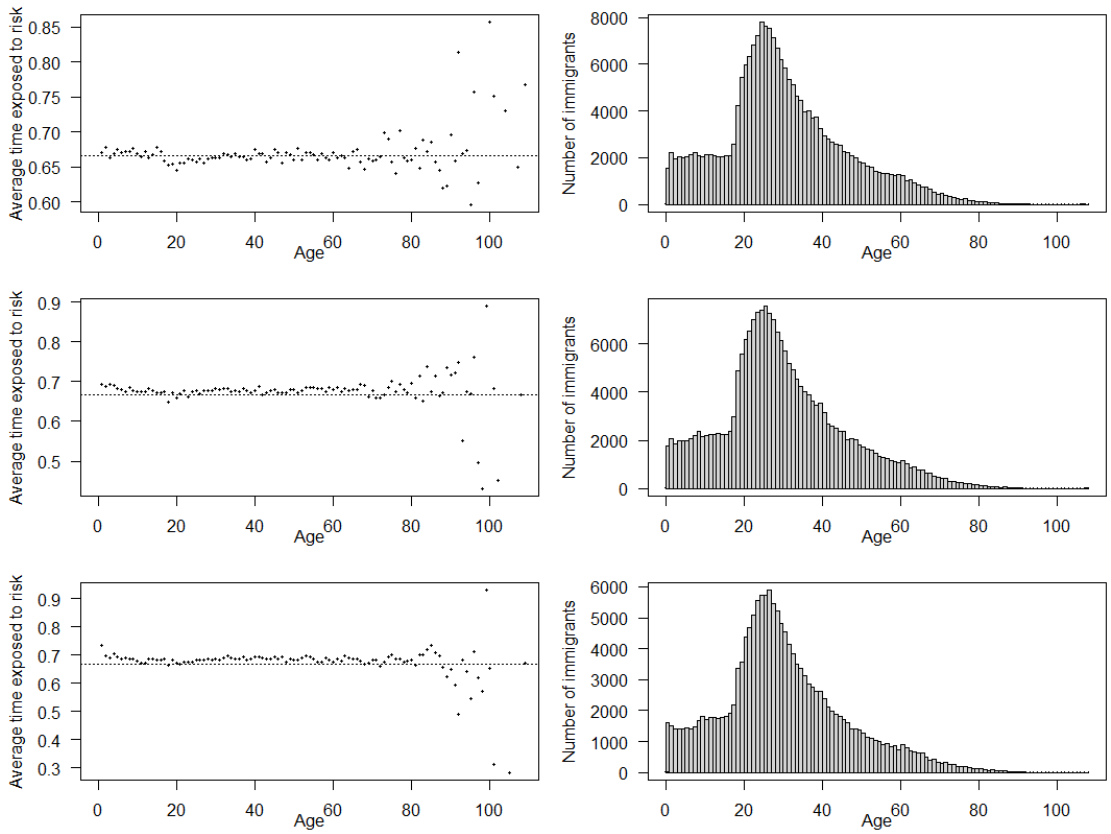

Figure 35A. Average number of years exposed to risk and number of immigrants by age for females immigrating in lower triangles in years 2006 (upper), 2007 (middle) and 2008 (lower).

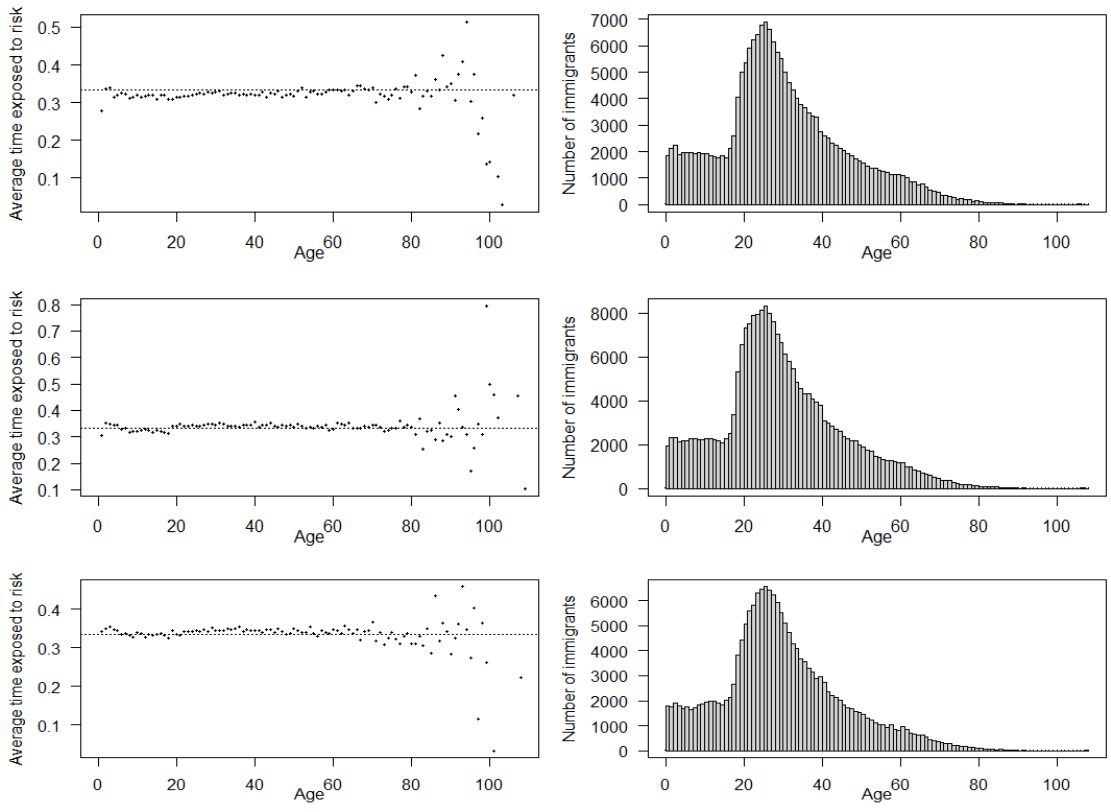

Figure 36A. Average number of years exposed to risk and number of immigrants by age for females immigrating in upper triangles in years 2006 (upper), 2007 (middle) and 2008 (lower).

LIFE TABLE FIGURES

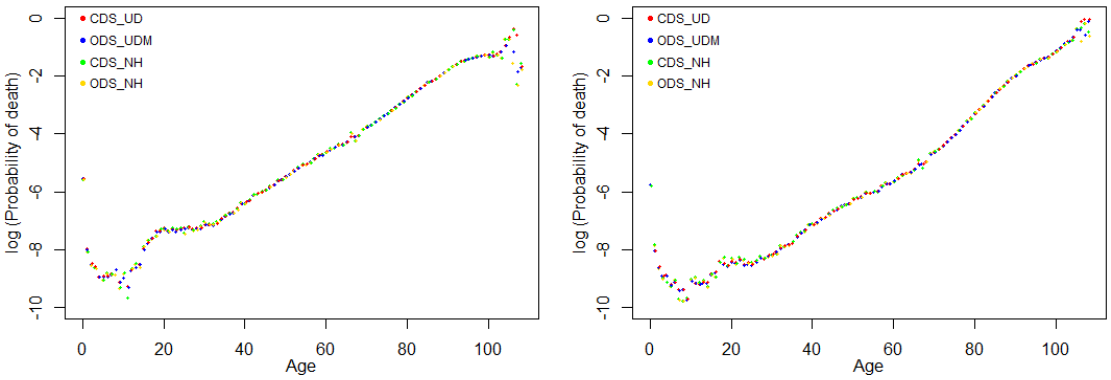

Figure 37A. Crude 2006-2007 cohort-based estimated life tables for men (left panel) and women (right panel). Graphical representation, in logarithmic scale, of the estimated crude probabilities of death by age for the four different data availability scenarios: CDS_UD, closed demographic system and uniform distribution of deaths by age and calendar year; ODS_UDM, open demographic system and uniform distribution of deaths and migrants; CDS_NH, closed demographic system with no hypothesis about distribution of deaths; and, ODS_NH, open demographic system with no hypotheses about distribution of deaths and migrants.

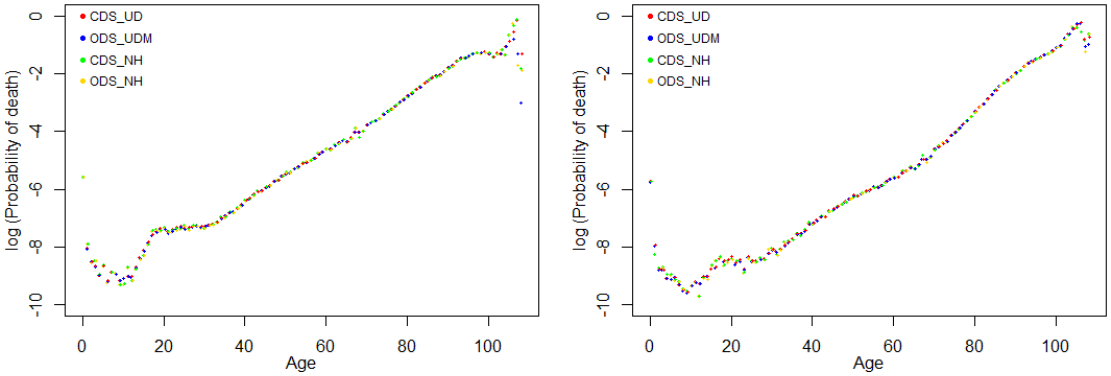

Figure 38A. Crude 2007-2008 cohort-based estimated life tables for men (left panel) and women (right panel). Graphical representation, in logarithmic scale, of the estimated crude probabilities of death by age for the four different data availability scenarios: CDS_UD, closed demographic system and uniform distribution of deaths by age and calendar year; ODS_UDM, open demographic system and uniform distribution of deaths and migrants; CDS_NH, closed demographic system with no hypothesis about distribution of deaths; and, ODS_NH, open demographic system with no hypotheses about distribution of deaths and migrants.

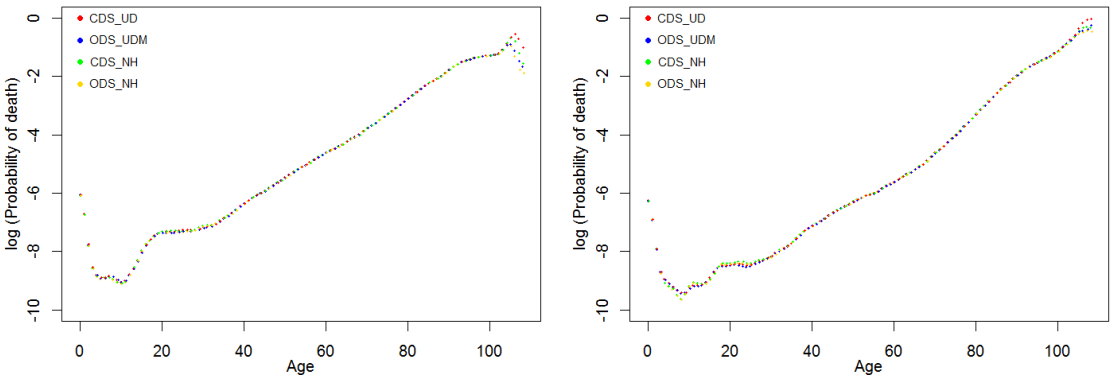

Figure 39A. Graduated 2006-2007 cohort-based estimated life tables for men (left panel) and women (right panel). Graphical representation, in logarithmic scale, of the graduated probabilities of death by age for the four different data availability scenarios: CDS_UD, closed demographic system and uniform distribution of deaths by age and calendar year; ODS_UDM, open demographic system and uniform distribution of deaths and migrants; CDS_NH, closed demographic system with no hypothesis about distribution of deaths; and, ODS_NH, open demographic system with no hypotheses about distribution of deaths and migrants.

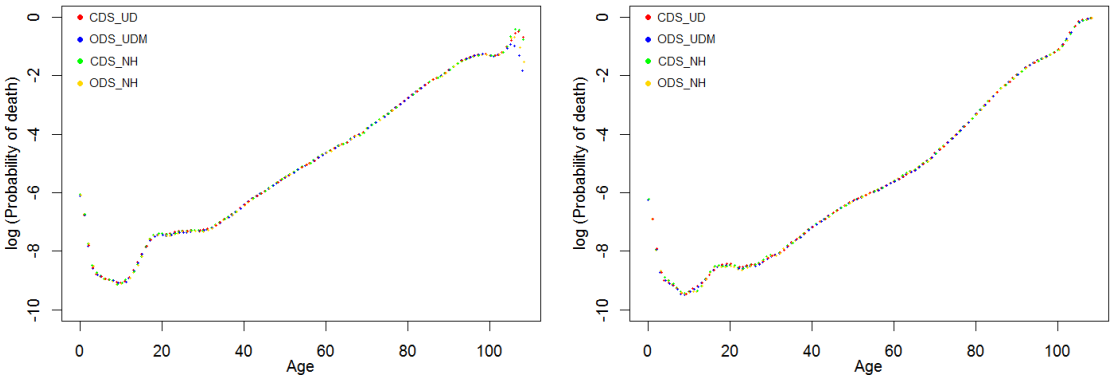

Figure 40A. Graduated 2007-2008 cohort-based estimated life tables for men (left panel) and women (right panel). Graphical representation, in logarithmic scale, of the graduated probabilities of death by age for the four different data availability scenarios: CDS_UD, closed demographic system and uniform distribution of deaths by age and calendar year; ODS_UDM, open demographic system and uniform distribution of deaths and migrants; CDS_NH, closed demographic system with no hypothesis about distribution of deaths; and, ODS_NH, open demographic system with no hypotheses about distribution of deaths and migrants.

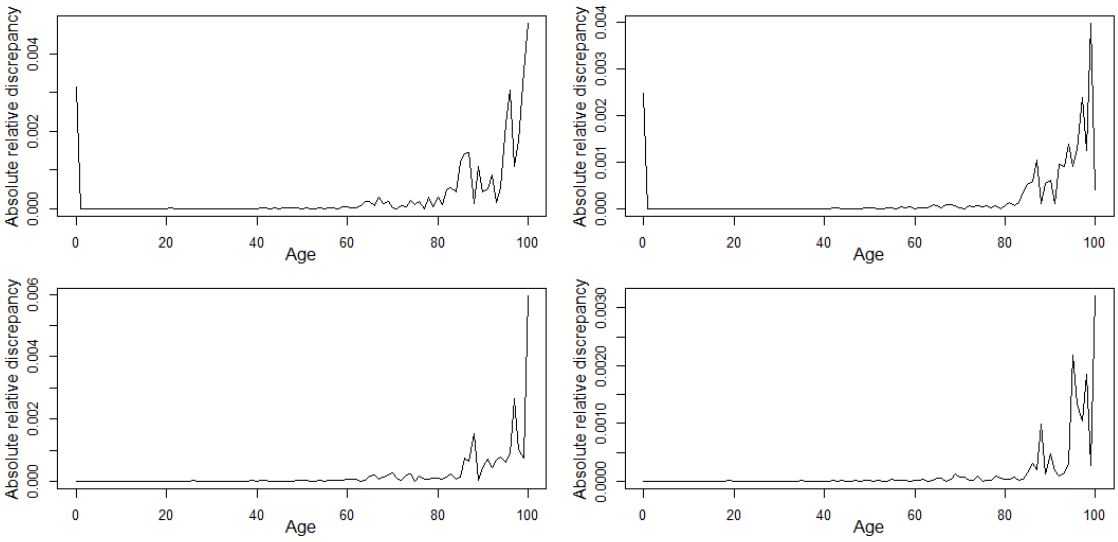

Figure 41A. Absolute relative discrepancies between the crude estimated probabilities of death obtained directly and indirectly via m_x under the closed demographic system and uniform distribution of deaths by age and calendar year (CDS_UD) for 2006-2007 men (upper left), 2006-2007 women (upper right), 2007-2008 men (lower left) and 2007-2008 women (lower right).

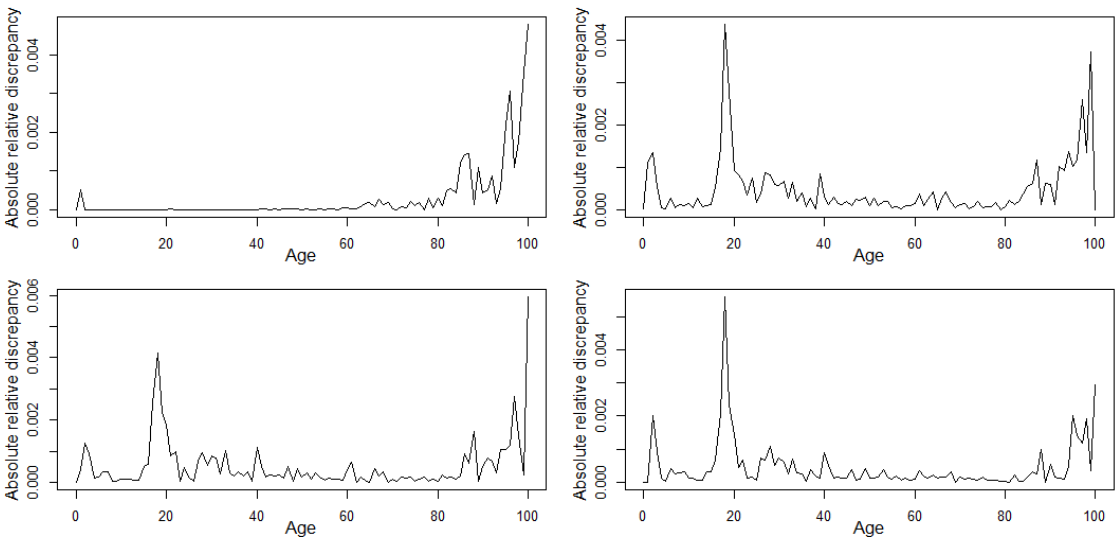

Figure 42A. Absolute relative discrepancies between the crude estimated probabilities of death obtained directly and indirectly via m_x under the open demographic system and uniform distribution of deaths and migrants (ODS_UDM) scenario for 2006-2007 men (upper left), 2006-2007 women (upper right), 2007-2008 men (lower left) and 2007-2008 women (lower right).

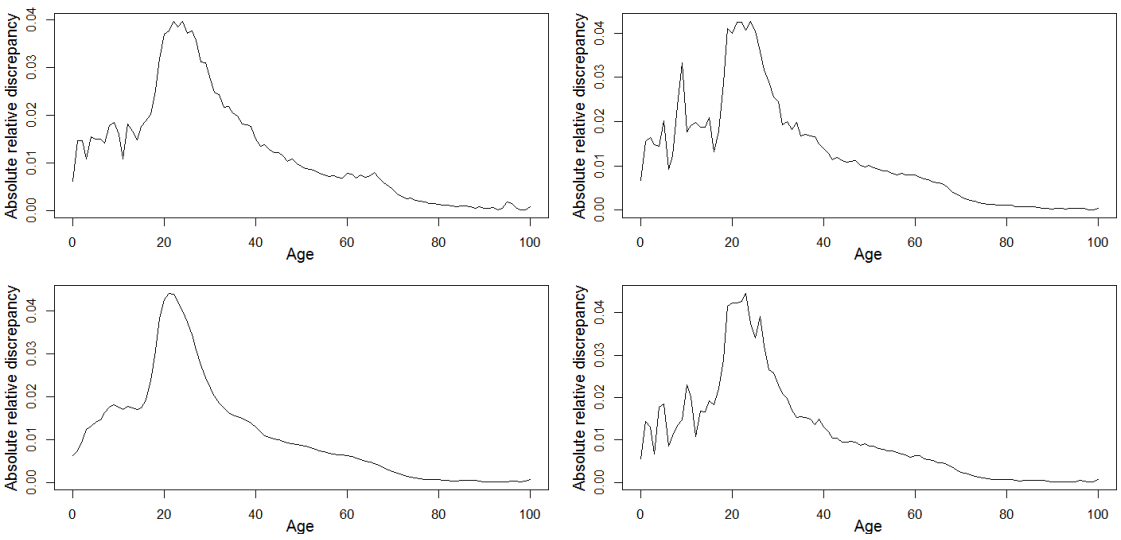

Figure 43A. Absolute relative discrepancies, $\frac{|\tilde{q}_x - \hat{q}_x|}{\hat{q}_x}$, between the crude estimated probabilities of death obtained under the closed demographic system and uniform distribution of deaths by age and calendar year (CDS_UD) scenario and the open demographic system and uniform distribution of deaths and migrants (ODS_UDM) scenario for 2006-2007 men (upper left), 2006-2007 women (upper right), 2007-2008 men (lower left) and 2007-2008 women (lower right).

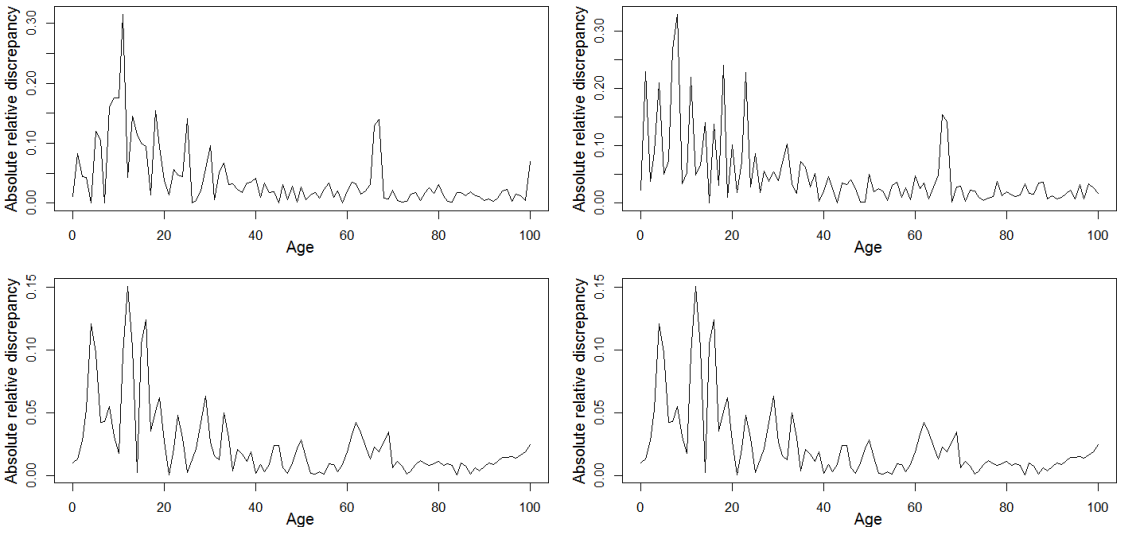

Figure 44A. Absolute relative discrepancies, $\frac{|\tilde{q}_x - \hat{q}_x|}{\hat{q}_x}$, between the crude estimated probabilities of death obtained under the closed demographic system and uniform distribution of deaths by age and calendar year (CDS_UD) scenario and the closed demographic system with no hypothesis about distribution of deaths scenario (CDS_NH) scenario for 2006-2007 men (upper left), 2006-2007 women (upper right), 2007-2008 men (lower left) and 2007-2008 women (lower right).

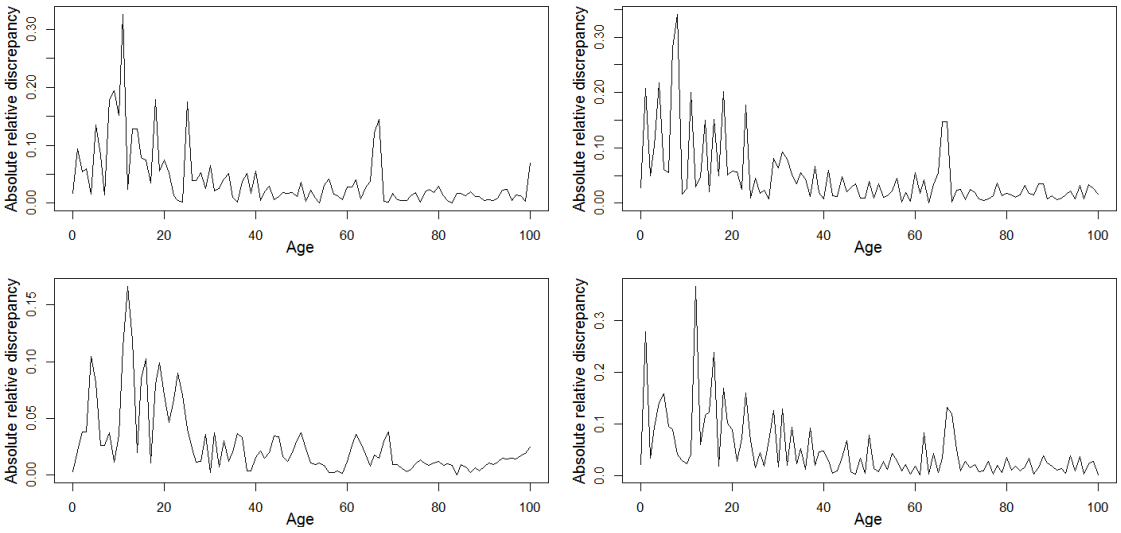

Figure 45A. Absolute relative discrepancies, $\frac{|\ddot{q}_x - \hat{q}_x|}{\hat{q}_x}$, between the crude estimated probabilities of death obtained under the closed demographic system and uniform distribution of deaths by age and calendar year (CDS_UD) scenario and the open demographic system with no hypotheses about distribution of deaths and migrants scenario (ODS_NH) scenario for 2006-2007 men (upper left), 2006-2007 women (upper right), 2007-2008 men (lower left) and 2007-2008 women (lower right).

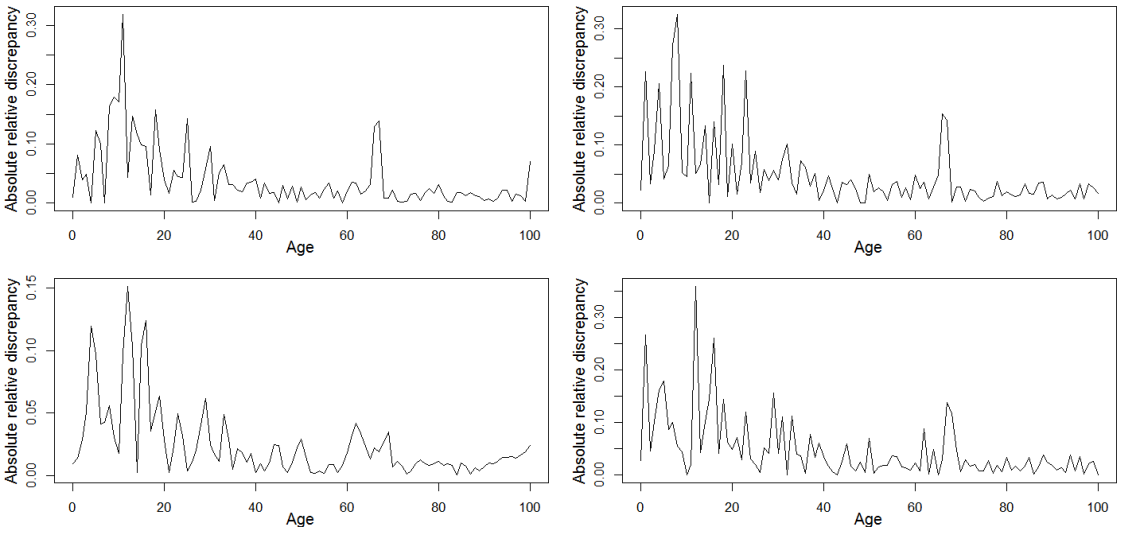

Figure 46A. Absolute relative discrepancies, $\frac{|\ddot{q}_x - \tilde{q}_x|}{\tilde{q}_x}$, between the crude estimated probabilities of death obtained under the open demographic system and uniform distribution of deaths and migrants (ODS_UDM) scenario and the open demographic system with no hypotheses about distribution of deaths and migrants scenario (ODS_NH) scenario for 2006-2007 men (upper left), 2006-2007 women (upper right), 2007-2008 men (lower left) and 2007-2008 women (lower right).

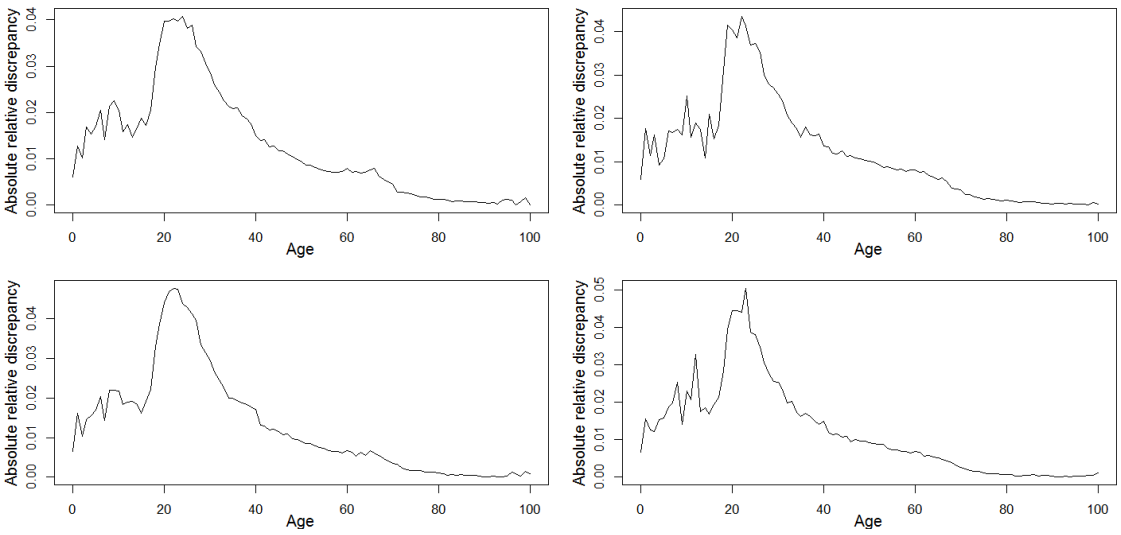

Figure 47A. Absolute relative discrepancies, $\frac{|\ddot{q}_x - \breve{q}_x|}{\breve{q}_x}$, between the crude estimated probabilities of death obtained under scenario and the closed demographic system with no hypothesis about distribution of deaths scenario (CDS_NH) scenario and the open demographic system with no hypotheses about distribution of deaths and migrants scenario (ODS_NH) scenario for 2006-2007 men (upper left), 2006-2007 women (upper right), 2007-2008 men (lower left) and 2007-2008 women (lower right).

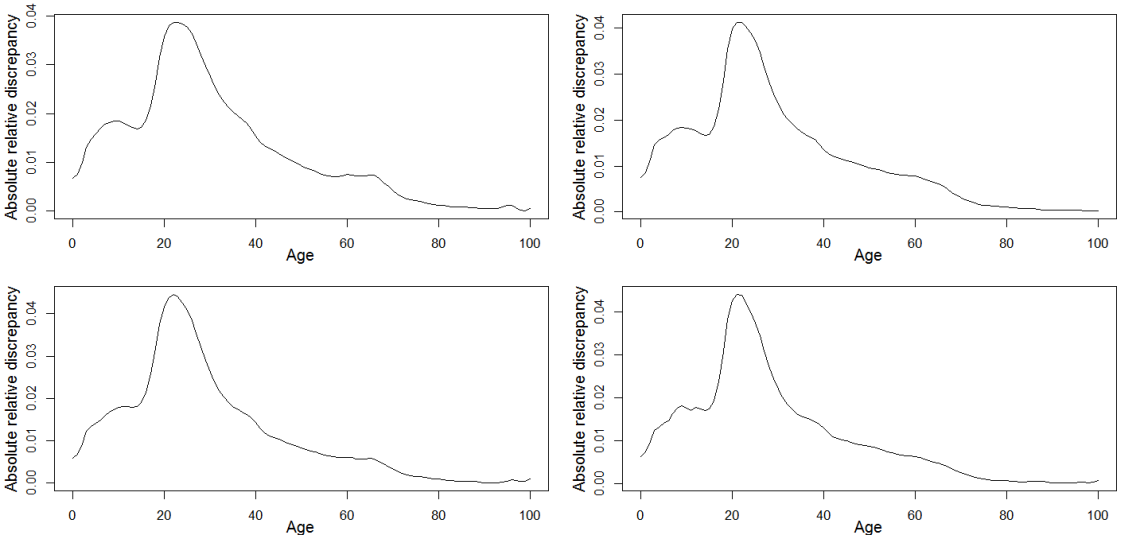

Figure 48A. Absolute relative discrepancies, $\frac{|\breve{q}_x - \hat{q}_x|}{\hat{q}_x}$, between the graduated probabilities of death obtained under the closed demographic system and uniform distribution of deaths by age and calendar year (CDS_UD) scenario and the open demographic system and uniform distribution of deaths and migrants (ODS_UDM) scenario for 2006-2007 men (upper left), 2006-2007 women (upper right), 2007-2008 men (lower left) and 2007-2008 women (lower right).

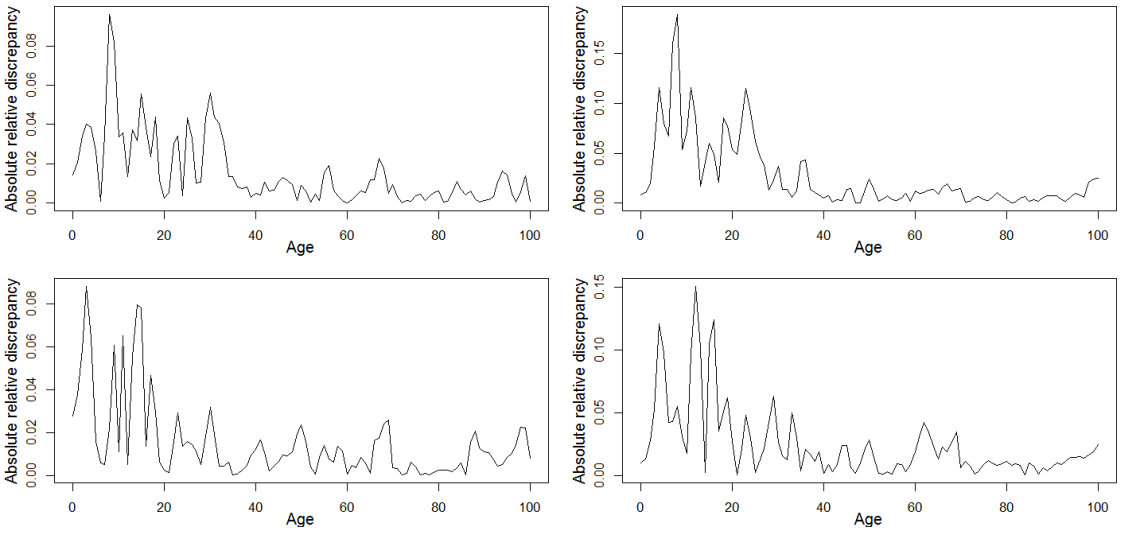

Figure 49A. Absolute relative discrepancies, $\frac{|\ddot{q}_x - \hat{q}_x|}{\hat{q}_x}$, between the graduated probabilities of death obtained under the closed demographic system and uniform distribution of deaths by age and calendar year (CDS_UD) scenario and the closed demographic system with no hypothesis about distribution of deaths scenario (CDS_NH) scenario for 2006-2007 men (upper left), 2006-2007 women (upper right), 2007-2008 men (lower left) and 2007-2008 women (lower right).

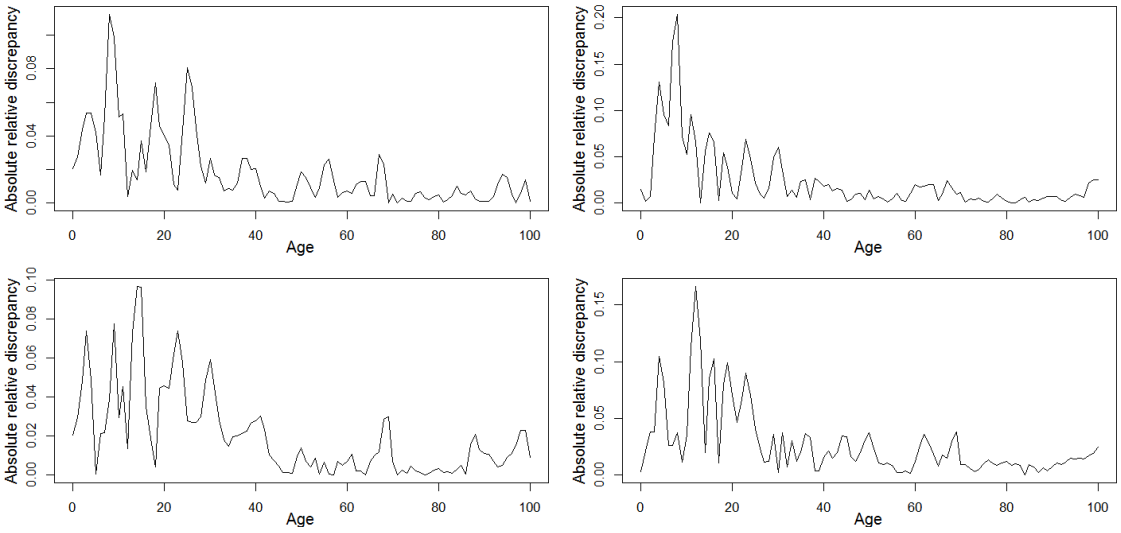

Figure 50A. Absolute relative discrepancies, $\frac{|\ddot{q}_x - \hat{q}_x|}{\hat{q}_x}$, between the graduated probabilities of death obtained under the closed demographic system and uniform distribution of deaths by age and calendar year (CDS_UD) scenario and the open demographic system with no hypotheses about distribution of deaths and migrants scenario (ODS_NH) scenario for 2006-2007 men (upper left), 2006-2007 women (upper right), 2007-2008 men (lower left) and 2007-2008 women (lower right).

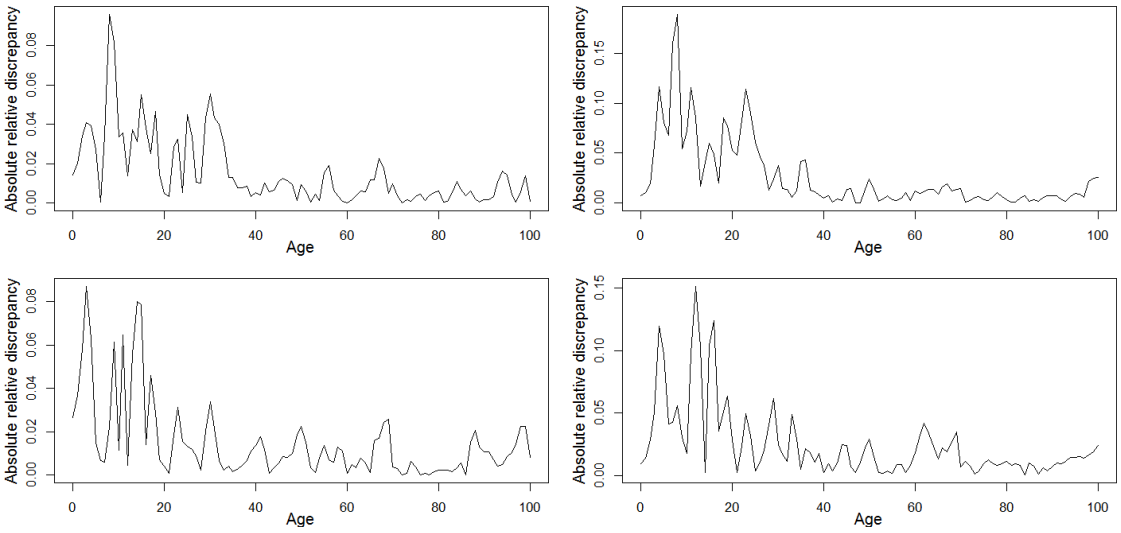

Figure 51A. Absolute relative discrepancies, $\frac{|\ddot{q}_x - \tilde{q}_x|}{\tilde{q}_x}$, between the graduated probabilities of death obtained under the open demographic system and uniform distribution of deaths and migrants (ODS_UDM) scenario and the open demographic system with no hypotheses about distribution of deaths and migrants scenario (ODS_NH) scenario for 2006-2007 men (upper left), 2006-2007 women (upper right), 2007-2008 men (lower left) and 2007-2008 women (lower right).

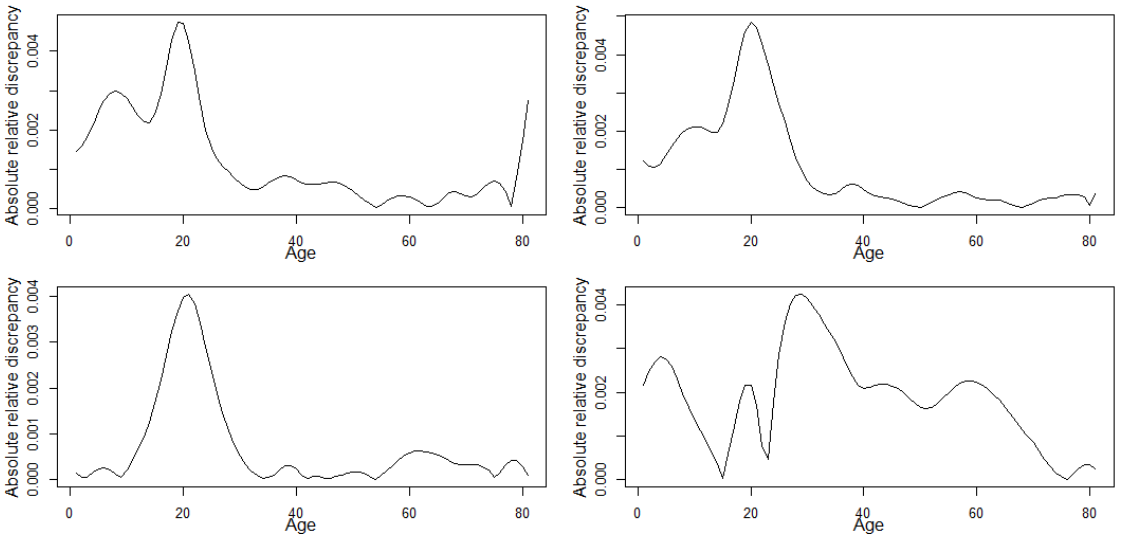

Figure 52A. Absolute relative discrepancies, $\frac{\left|\ddot{q}_x^{AB} - \ddot{q}_x^{FC}\right|}{\ddot{q}_x^{AB}}$, between the graduated probabilities of death obtained under the open demographic system and uniform distribution of deaths and migrants (ODS_UDM) scenario for the ages with relevant migration flows with BC computed, respectively, from either AB or FC for 2006-2007 men (upper left), 2006-2007 women (upper right), 2007-2008 men (lower left) and 2007-2008 women (lower right).

Incorporating big microdata in life table construction: A hypothesis-free estimator (A3)

Josep Lledó, Universitat de Valencia

Jose M. Pavía, Universitat de Valencia

Francisco G. Morillas, Universitat de Valencia

Abstract

The IT revolution, now more than ever, offers a cheaper and faster way to collect, store, transmit and process data. Detailed microdata of dates of death, migration and birth are already available for general populations. In this paper, we develop within the family of period-based estimators a new, assumption-free estimator for constructing life tables. The estimator proposed exploits all the detailed data available and is free of the theoretical inconsistencies that the estimators currently used by most official statistical agencies have. We compute the proposed estimator for a real database and test the suitability of the hypotheses on which the estimators used so far rely. The hypothesis of uniform distribution of birthdays is proven to be inadequate and the one having the largest impact on the estimated probabilities. Given its influence on public pension systems and life insurances, we advocate for adopting the more efficient approaches proposed in this paper.

Keywords: rates of mortality; date of births; period-based estimators; big data; Spain.

Acknowledgements

The authors wish to thank *Instituto Valenciano de Estadística*, mainly Francisco Fabuel, for their first-rate assistance in providing the detailed statistics of birth dates handled in this research. Our thanks are also due to M. Hodkinson for translation of the paper into English. This work has been supported by the Spanish Ministry of Economics and Competitiveness under grants CSO2013-43054-R and MTM2016-74921-P.

1. Introduction

The measurement of mortality is a topical issue which has received much attention in the demographic-actuarial literature throughout the 20th century (e.g., Forsyth, 1914; Coale and Demey, 1966; Benjamin and Pollard, 1986; Preston *et al.*, 2001). Over the last few years, the focus has been more on the analysis of its evolution and the study of longevity risk. Both these areas of scrutiny are necessary for policymaking and public planning (for example, for predicting future financial requirements in social security public systems) and in life assurance and pensions products (where pricing and reserving operations are crucial for life insurers and pension plans to properly manage risk exposures). In this regard, numerous publications have addressed different aspects of the problem (e.g., Lee and Carter, 1992; Lee and Miller, 2001; Tabeau *et al.*, 2001; Blake *et al.*, 2006; Biffis and Blake, 2009; Pitacco *et al.*, 2009; Cairns *et al.*, 2011; Li and Hardy, 2011; Barrieu *et al.*, 2012; Börger *et al.*, 2014; or Enchev *et al.*, 2016).

In recent years, we have also witnessed a major revolution in information technology, a revolution which is having a significant impact on society and business, including the insurance business (Cummins and Santomero, 2005; IABE, 2015). According to Ruggles (2014), this revolution could also transform the demographic-actuarial research, for example, by incorporating big microdata relating to immigration, emigration or birth events into the demographic computations (e.g., of mortality). Up until now, these data have rarely been taken into account despite it being proved that, for instance, migratory flows can have a measurable and not negligible impact on the life table (Lledó *et al.*, 2016).

The life table (or mortality table) is the tool most used to synthesize the mortality of a collective. Among other biometric features, a life table collects the probabilities of the members of a particular population surviving to, p_x, or dying within one year, $q_x = 1 - p_x$, at each integral age x. General population life tables are constructed after estimating either mortality rates, m_x, or death probabilities, q_x. To estimate m_x or q_x it is common to use estimators that, depending on the level of detail of the information available, assume certain hypotheses implicit in their construction. When working with period-based estimators (e.g., Wilmoth *et al.*, 2007; ONS, 2010, 2012; INE, 2009, 2016; Arias, 2015), we find, among the usual implicit hypotheses, the assumptions of: (H1)

uniform distribution of deaths (and migrants) for each age and calendar year, (H2) closed demographic system (or at least no explicit consideration of migration flows), and (H3) uniform distribution of birthdays of individuals who survived the year. The above hypotheses, however, are not innocuous. For instance, Pavía *et al.* (2012) have verified, using bi-annual cohort-based estimators, that the hypothesis (H2) entails a significant underestimation of the probabilities of death in a scenario in which immigration flows are much higher than emigration flows. Despite this, over recent years, official statistical agencies have moved towards a progressive replacement of the reference estimators: from the less assumption-demanding *q*-type (cohort-based) estimators (e.g., INE, 2007; Arias *et al.*, 2010) to the most assumption-demanding *m*-type (period-based) estimators (e.g., ONS, 2012, Arias, 2015, INE, 2016). The hypothesis (H3) is unnecessary when one works with *cohort-based estimators*.

The aim of this paper is twofold. On the one hand, within the family of *period-based estimators*, we propose an estimator for m_x free from all the previous hypotheses. This is especially relevant because, as shown in the statistical annex at the foot of this article, the above hypotheses may lead to different analytical solutions depending on the reasoning followed, something that does not happen within the same conceptual framework with the *q*-type estimators. On the other hand, we evaluate, using a real database, the set of assumptions (H1)–(H3). To assess hypotheses and their impact, further calculations of rates and probabilities of death are carried out using different estimators that vary in their requirements for detailed information. As a dataset, we use the microdata files corresponding to the years 2010 to 2013 of the Comunitat Valenciana (Spain).

At this point and before continuing, we want to highlight two issues. Firstly, although we just consider one-year *period-based estimators* (Wilmoth *et al.*, 2007; INE, 2009, 2016; Arias, 2015), we should note that it is equally as simple to generalise the use of our estimator for dealing with multi-year *period-based estimators* (ONS, 2010, 2012; Arias, 2015; ABS, 2016). And, secondly, although there are different approaches in the literature for estimating the denominator of m_x, we use the total number of 'person-years' at risk. Compared to the approach of using the average population at risk of dying (measured by mid-year population estimates), the total number of 'person-

years' at risk has the advantage of establishing a closer correspondence between deaths and exposed-to-risk: all deaths are members of the exposed-to-risk population. However, when mid-year population estimates are used (e.g., ONS, 2012; Arias, 2015), the numerator of m_x could more than likely account for deaths that have not even been considered in the population at risk.

The rest of the work is structured as follows. The second section presents the terminology and methodology used and introduces the formulas of the five estimators, which collapse into four, that are studied in this research. Each one of the considered estimators corresponds to a scenario with different requirements of detailed information. Section three introduces the statistical hypothesis tests implemented to analyse the appropriateness of the hypotheses (H1)–(H3). The results of the application of the tests is presented in section four. Section five analyses the differences that would arise from using each of the estimators introduced in section two and examines their impact on some insurance products. The sixth section discusses and summarizes the results. Finally, in the annex, we show the theoretical weaknesses of the assumptions (H1)–(H3). Supplementary (online) material complements the paper.

2. Preliminaries and methodology

A period life table shows the mortality experience of a hypothetical cohort of new born babies, based on the assumption that during the course of their lifetime the group faces the age-specific mortality rates of the reference period. There is a long tradition of analysis of mortality through life tables. The first attempts at their construction began in the 17th century. In 1662, Graunt presented the first estimates of death rates by analysing data of deaths in London (Graunt, 1662). In 1963, Halley published the table of mortality of Breslau, which was used by the British government to sell life annuities with the prices adjusted depending on the age of the purchasers. In 1746, Depardieux estimated the mortality table of the French population and, around 1770, Cambert calculated the mortality table of the German population (Basulto and Garcia, 2009). The first conceptualizations of the phenomenon through models, however, came later: Gompertz (1825), Makeham (1860).

The introduction of graphic representation as an auxiliary tool and the conceptualization and estimation of mortality appeared in parallel. The so-called Lexis scheme dated from 1870 (Brasche, 1870; Lexis, 1880; Vandesrchirk, 2001). The Lexis scheme is used to graphically represent the behaviour of different demographic events that affect a population at a specific time. In the present work, we derive the mortality rate estimators, m_x, based on the Lexis scheme, constructing the life table as a function of the available information. In this sense, the information corresponding to raw data (micro data) relating to deaths, migratory flows and dates of birth is used. In order to facilitate understanding of the variables used, Table 1 presents the notation.

Table 1. Detail of symbols used in the equations.

Notation	Description
x	Age, measured in years.
t	Calendar year, measured in years.
q_x	Probability of death during the age interval x to $x+1$.
m_x	(Central) rate of mortality at age x.
C_x^t	(Census) population with completed age (age at last birthday) x on January 1 of year t.
$C_{x,d}^t$	(Census) population with exact age x years and d days on January 1 of year t.
D_x^t	Number of deaths in year t with completed age x.
$D_{x:L}^t$	Number of deaths in year t with completed age x, born in year $x-t$.
$D_{x:U}^t$	Number of deaths in year t with completed age x, born in year $x-t-1$.
E_x^t	Number of emigrants in year t with completed age x.
$E_{x:L}^t$	Number of emigrants in year t with completed age x, born in year $x-t$.
$E_{x:U}^t$	Number of emigrants in year t with completed age x, born in year $x-t-1$.
I_x^t	Number of immigrants in year t with completed age x.
$I_{x:L}^t$	Number of immigrants in year t with completed age x, born in year $x-t$.
$I_{x:U}^t$	Number of immigrants in year t with completed age x, born in year $x-t-1$.
$d_{x,j}^t$	For the jth deceased person of year t with completed age x, the minimum of the difference in years between the moment of their death and either the date of their x birthday or January 1 of year t.
$e_{x,j}^t$	For the jth emigrant of year t with completed age x, the minimum of the difference in years between their moment of emigration and either the date of their x birthday or January 1 of year t.
$i_{x,j}^t$	For the jth immigrant of year t with completed age x, the minimum of the difference in years between their moment of immigration and either the date of their $x+1$ birthday or January 1 of year $t+1$.
$l_{x,j}^t$	For the jth migrant or deceased person of year t with completed age x, the years lived by the person with age x (i.e., since their x birthday) up to the moment of the migrant or death event.

The Lexis scheme (Figure 1) is a two-dimensional diagram, of age x and calendar year or period t, which shows the personal history of each of the individuals involved, represented by a linear segment of unit slope. Each personal history begins with birth (lines beginning at the base of the diagram) or with the event of immigration, denoted by ´o´ (example: line b in Figure 1-left) and ends with death, denoted by ´x´ (example: line a in Figure 1-left), or with emigration, denoted by '□' (example; line c in Figure 1-

left). People who continue to live at age x at the end of year t are also represented (example; line d in Figure 1-left).

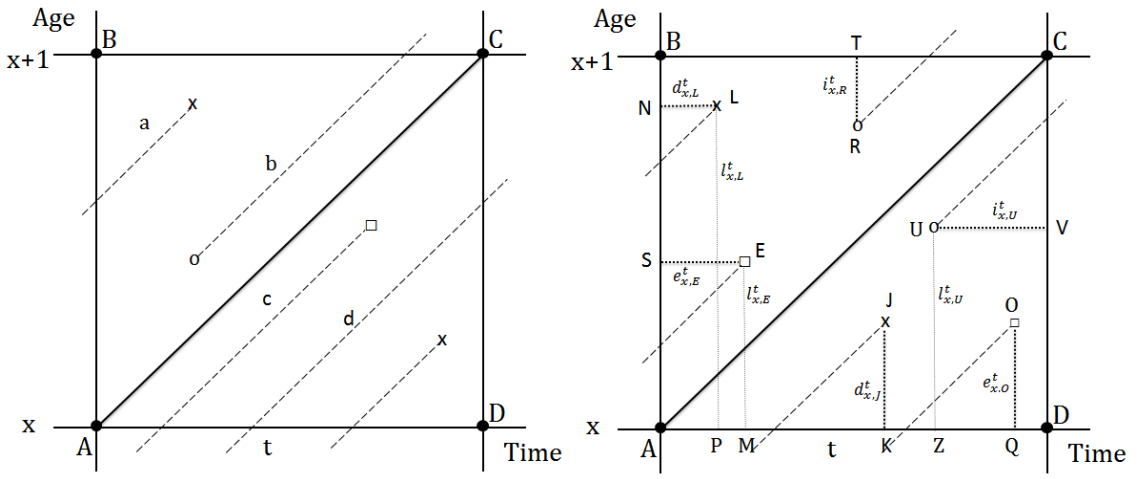

Figure 1. Left panel: Small section of a 1x1 cell of the Lexis diagram with some lifelines and a schematic representation of death (×) and migrant events (immigrant ○, emigrant □). Right panel: Detail of a 1x1 cell of the Lexis diagram with some death and migrant events. $d^t_{x,J}$ and $e^t_{x,O}$ measure the distance in years between either the events of death J or emigration O and the dates of their x birthdays. $d^t_{x,L}$ and $e^t_{x,E}$ measure the distance between either the events of death L or emigration E and January 1 of year t. $i^t_{x,R}$ and $i^t_{x,U}$ measure, respectively, the distances between the events of immigration of U and R and January 1 of year $t+1$ or the dates of their $x+1$ birthdays, whichever happens first. Note that except in exceptional cases (such as when a person immigrates and dies, immigrates and emigrates or emigrates and immigrates with age x in year *t*), the above quantities account for their exact time (in years) exposed to the risk of dying during year t with completed age x as a member(s) of the target population. Finally, $l^t_{x,L}$, $l^t_{x,E}$ and $l^t_{x,U}$ measure the distance in years between either the events of death L or migration E and U and the dates of their x birthdays. They represent the years lived by the person with age x up to the moment of the event. Observe that if we extended this last notation to the rest of points we would have $d^t_{x,J} = l^t_{x,J}$, $e^t_{x,O} = l^t_{x,O}$ and $i^t_{x,R} = 1 - l^t_{x,R}$. Point-events and individuals are identified with the same symbol to alleviate notation.

Another use of the Lexis scheme is to show, over calendar time and age, the stocks and flows of the population by assigning values to its segments and surfaces. In Figure 1, the length of segment AB (DC) represents the number of people in the population living with age x as of January 1 of year t ($t+1$), denoted by C^t_x (C^{t+1}_x). The number of deaths occurring at age x in year *t* corresponds to the total number of lifelines ending at a cross on the ABCD surface, denoted by D^t_x. In turn, we denote $D^t_{x:L}$ as the number of people born in year $x-t$ who have passed away with age x throughout the year t (crosses ending in the lower triangle ACD) and as $D^t_{x:U}$ the number of people born in year $x-t-1$ who have passed away with age x along t (crosses ending in the upper triangle ABC).

With respect to migratory flows, the number of squares (circles) ending (starting) at the surface ABCD corresponds to the number of emigrants (immigrants) emigrating (immigrating) with age x in year t, denoted by E_x^t (I_x^t). In turn, $E_{x:L}^t$ ($I_{x:L}^t$) denotes the number of people born in year $x-t$ who have emigrated (immigrated) reaching age x during year t (squares that end, circles that begin, in the lower triangle ACD). On the other hand, $E_{x:U}^t$ ($I_{x:U}^t$) corresponds to the number of people born in year $x-t-1$ who have emigrated (immigrated) reaching age x during year t (squares starting, circles ending, in the upper triangle ABC).

When detailed information is available on dates of demographic events and such information is deemed relevant for incorporation into the estimation of mortality rates, new variables should be introduced into the Lexis scheme (see Figure 1-right). $d_{x,j}^t$ and $e_{x,O}^t$ (segments JK and OQ in Figure 1-right) are examples of the elapsed time (in years) between the date of death/emigration and birthday date at age x for individuals who died/emigrated with age x and reached that age during year t (crosses or squares ending in the lower triangle ACD). Assuming that they did not immigrate to the collective under study during year t, these distances represent the time in years that they were at risk-of-dying during year t with completed age x as a member of the target population. In the opposite case, the lengths of the segments NL and SE, $d_{x,L}^t$ and $e_{x,E}^t$, are examples of times exposed to risk for those individuals who died/emigrated with age x in year t but who reached the age x during year $t-1$ (crosses or squares that finish in the upper triangle, ABC).

In the case of immigrants born in year $x-t$ (circles beginning in the lower triangle ACD in Figure 1, for example U), $i_{x,U}^t$ is defined as the difference in years between the date of immigration and January 1 of year $t+1$ (length of the UV segment), while for immigrants who fall into the study group in the upper triangle ABC, for example R, $i_{x,R}^t$ is defined as the difference in years between their immigration date and the date of their next birthday (RT segment length). Assuming that they do not die during year t, these figures represent the time in years that they are at risk of dying during year t with completed age x as a member of the population under study. Finally, the lengths of the segments LP, EM and UZ, denoted by $l_{x,L}^t$, $l_{x,E}^t$ and $l_{x,U}^t$, measure the time in years lived by L, E and U since their x birthdays up to either their corresponding events of either

death, L, or migration, E and U, occur. Note that their complements to one ($1 - \mathrm{l}^{\mathrm{t}}_{\mathrm{x,L}}$, $1 - \mathrm{l}^{\mathrm{t}}_{\mathrm{x,E}}$ and $1 - \mathrm{l}^{\mathrm{t}}_{\mathrm{x,U}}$) are the times that they have not been at risk of dying as a member of the target population with completed age x in year t.

To estimate mortality rates, for each age group (and sex) we need to know the total number of deaths and total time exposed to risk of dying of the population under study during the period of analysis. The exact way in which the time exposed to risk is calculated will depend on the level of detail available in the data and, as we shall see in the annex, the reasoning followed. In this research, as a starting point to compute the total number of 'person-years' at risk, we do not start from theoretical demographic entities that are usually represented in the Lexis scheme (such as the total number of people reaching the exact age x throughout year t, segment AD), but rather by considering the type of data that are usually produced by official statistical systems, i.e., stocks of populations and flows of migrants and deceased. Hence, under the hypothesis of uniform distribution of dates of birth, we attain an initial estimate of the total number of exposed-to-risk as the average of the populations registered with completed age x at January 1 of year t, C_x^t, and January 1 of year $t+1$, C_x^{t+1} (segments AB and CD, Figure 1). Afterwards, this initial estimate is adjusted excluding/including the time exposed to risk of people who die, immigrate or emigrate with age x during year t. As a way of summarising what will be explained and justified in the following subsections, Table 2 shows the one-year period-based estimator for m_x that we will derive under different hypotheses.

Table 2. Summary of estimators and hypotheses used for derivation.

Estimator	Hypotheses			
	Closed population (CP)	Uniform Deaths (D)	Uniform Migrants (M)	Uniform Dates of Birth (B)
$\widehat{m}_x = \frac{D_x^t}{\frac{1}{2}C_x^t - \frac{1}{6}D_x^t + \frac{1}{2}C_x^{t+1} + \frac{1}{6}D_x^t} = \frac{D_x^t}{\frac{1}{2}C_x^t + \frac{1}{2}C_x^{t+1}}$	Yes	Yes	N/A	Yes
$\bar{m}_x = \frac{D_x^t}{\frac{1}{2}C_x^t - \frac{1}{6}D_x^t - \frac{1}{6}E_x^t + \frac{1}{6}I_x^t + \frac{1}{2}C_x^{t+1} + \frac{1}{6}D_x^t + \frac{1}{6}E_x^t - \frac{1}{6}I_x^t} = \frac{D_x^t}{\frac{1}{2}C_x^t + \frac{1}{2}C_x^{t+1}}$	No	Yes	Yes	Yes
$\tilde{m}_x = \frac{D_x^t}{\frac{1}{2}C_x^t - \sum_{j=1}^{D_{x:U}^t}(1 - l_{x,j}^t) + \frac{1}{2}C_x^{t+1} + \sum_{j=1}^{D_{x:L}^t} d_{x,j}^t}$	Yes	No	N/A	Yes
$\ddot{\mathrm{m}}_x = \frac{D_x^t}{\frac{1}{2}C_x^t - \sum_{j=1}^{D_{x:U}^t}(1 - l_{x,j}^t) - \sum_{j=1}^{E_{x:U}^t}(1 - l_{x,j}^t) + \sum_{j=1}^{I_{x:U}^t} i_{x,j}^t + \frac{1}{2}C_x^{t+1} + \sum_{j=1}^{D_{x:L}^t} d_{x,j}^t + \sum_{j=1}^{E_{x:L}^t} e_{x,j}^t - \sum_{j=1}^{I_{x:L}^t}(1 - l_{x,j}^t)}$	No	No	No	Yes
$\dddot{\mathrm{m}}_x = \frac{D_x^t}{\sum_{d=1}^{T}\frac{(d-0.5)}{T}C_{x,d}^t - \sum_{j=1}^{D_{x:U}^t}(1 - l_{x,j}^t) - \sum_{j=1}^{E_{x:U}^t}(1 - l_{x,j}^t) + \sum_{j=1}^{I_{x:U}^t} i_{x,j}^t + \sum_{d=1}^{T}\frac{(T-d+0.5)}{T}C_{x,d}^{t+1} + \sum_{j=1}^{D_{x:L}^t} d_{x,j}^t + \sum_{j=1}^{E_{x:L}^t} e_{x,j}^t - \sum_{j=1}^{I_{x:L}^t} l_{x,j}^t}$	No	No	No	No

N/A: Not applicable.

2.1. Closed population and uniform distribution of deaths and births (CP_UD_UB)

The scenario which requires the lowest level of information happens when migration flows are omitted and aggregate data of deaths and dates of birth are used. Under the hypothesis of even distribution of birthday dates, individuals counted in C_x^t (segment AB, Figure 1) will be, on average, ½ year of their lifetime during the year *t* at risk of dying with age *x*. With the same criterion, those who stayed alive to year *t+1*, C_x^{t+1} (segment CD, Figure 1), will contribute on average ½ year to the exposed-to-risk time with age x in the period *t*.

Likewise, under the hypothesis of uniform distribution of the deceased for each age and year, the number of deaths in the upper triangle (ABC) and in the lower triangle (ACD) will coincide and will be equal to half of the deceased during the whole year t, D_x^t. The deceased of age x in the period *t* that reached age x in the previous year (crosses that terminate their lifeline in the upper triangle, ABC) have been counted in C_x^t, and so it is necessary to subtract the time non-exposed to risk of these, $\frac{1}{2}D_x^t$, which on average is $\frac{1}{3}$ of the year (see, e.g., Willmoth *et al.*, 2007). Also, on average, we must add $\frac{1}{3}$ year of exposure for each deceased in the lower triangle (ACD). Hence, the estimator for the mortality rate, $\widehat{m}_x$, in this scenario is given by the equation (1).

$$\hat{m}_x = \frac{D_x^t}{\frac{1}{2}C_x^t - \frac{1}{6}D_x^t + \frac{1}{2}C_x^{t+1} + \frac{1}{6}D_x^t} = \frac{D_x^t}{\frac{1}{2}C_x^t + \frac{1}{2}C_x^{t+1}} \quad (1)$$

where $\hat{q}_x \approx \frac{\hat{m}_x}{1+\frac{1}{2}\hat{m}_x}$, assuming a constant force of mortality over the age interval.

2.2. Open population and uniform distribution of deaths, migrants and births (OP_UD_UM_UB)

When the closed population hypothesis is eliminated, migratory events should be taken into account and, under the hypothesis of uniformity, the migratory flows would be distributed evenly in the square ABCD, thus coinciding the number of immigrants and emigrants in both triangles: ABC and ACD. Likewise, under the same hypothesis, it is not difficult to prove that the average times exposed/non-exposed to risk of individuals migrating/immigrating in each triangle is on average $\frac{1}{3}$ of a year (Pavía *et al.*, 2012). Hence, in this scenario, the estimator for the mortality rate, $\bar{m}_x$, will be given by equation (2).

$$\bar{m}_x = \frac{D_x^t}{\frac{1}{2}C_x^t - \frac{1}{6}D_x^t - \frac{1}{6}E_x^t + \frac{1}{6}I_x^t + \frac{1}{2}C_x^{t+1} + \frac{1}{6}D_x^t + \frac{1}{6}E_x^t - \frac{1}{6}I_x^t} = \frac{D_x^t}{\frac{1}{2}C_x^t + \frac{1}{2}C_x^{t+1}} \quad (2)$$

Note that both the CP_UD_UB estimator, (1), and the ODS_UD_UM_UB estimator, (2), yield the same risk exposure time and, therefore, the same mortality rate.

2.3. Closed population with no hypotheses about distribution of deaths and uniform distribution of births (CP_NUD_UB)

Where a population is in its millions, the annual number of deaths is in its tens of thousands. For ease of handling such high numbers, the usual practice has been to assume, except for age zero, uniform distribution of deaths for each age and calendar year (e.g., INE, 2007; ONS, 2012; Ruggles, 2014; Arias, 2015). Several studies, however, have questioned the overall adequacy of such a hypothesis for certain ages and/or when extreme weather events, pandemics or armed conflicts occur (e.g., Deschenes *et al.*, 2007; Gavrilov and Gavrilova, 2011; Berko *et al.*, 2014; Lledó *et al.*, 2016) and have also verified that their impact on the insurance products is not innocuous (Fernández and Gregorio, 2015). Hence, such a hypothesis is now beginning to be abandoned (see, e.g.,

INE, 2016), because advances in computer engineering have made it possible to store and process a massive amount of data.

In the case of this scenario, the number of those exposed to risk with age x are obtained based on stock populations at January 1 of years t and $t+1$: C_x^t and C_x^{t+1}(Segments AB and CD). On the one hand, under the hypothesis of uniform distribution of birthday dates, each of the subjects who does not die during year *t* will be on average ½ year at risk of dying. On the other hand, for the dying subjects we know the exact time they remained members of the risk population with/without life with age x during year *t*. Hence, to calculate the total of 'person-years' at risk, (i) we deduce from $\frac{1}{2}C_x^t$ the total time that the deceased in ABC have not been at risk, i.e. $\sum_{j=1}^{D_{x:U}^t}(1-l_{x,j}^t)$, and (ii) we add to $\frac{1}{2}C_x^{t+1}$ the total time that the deceased in ACD have been at risk, i.e. $\sum_{j=1}^{D_{x:L}^t} d_{x,j}^t$. In summary, in this scenario, the estimator for the mortality rate, $\widetilde{m}_x$, is given by equation (3).

$$\widetilde{m}_x = \frac{D_x^t}{\frac{1}{2}C_x^t - \sum_{j=1}^{D_{x:U}^t}(1-l_{x,j}^t) + \frac{1}{2}C_x^{t+1} + \sum_{j=1}^{D_{x:L}^t} d_{x,j}^t} \qquad (3)$$

Finally, for the estimation of the corresponding probability of death, q_x, we need to compute the average number of years lived, f_x, by the deceased with age x throughout the year t that is: $\tilde{f}_x = \frac{\sum_j^{D_x^t} l_{x,j}^t}{D_x^t} = \frac{\sum_j^{D_{x:U}^t} l_{x,j}^t + \sum_j^{D_{x:L}^t} d_{x,j}^t}{D_{x:U}^t + D_{x:L}^t}$. From the previous expression, q_x is calculated using $\tilde{q}_x = \frac{\widetilde{m}_x}{1+(1-\tilde{f}_x)\widetilde{m}_x}$. The CP_UD_UB and the actual scenario are equal when the average time lived with age x by the deceased in year t,$\tilde{f}_x$, is ½.

2.4. Open population with no hypotheses about distribution of deaths and migrants and uniform distribution of births (OP_NUD_NUM_UB)

In demography, migratory flows play an important role in the evolution and development of any population (Khoo *et al.*, 2011). However, they are rarely explicitly considered in the construction of life tables in spite of the fact that some studies with cohort-based estimators have shown that, when the number of immigrants and

migrants are not equal, omitting migratory flows implies differences of up to 4% in the probability of death (Pavía *et al.*, 2012; Lledó *et al.*, 2016).

On the one hand, to obtain the total 'person-years' at risk with age x during year *t*, it is estimated that, under the assumption of an even distribution of birthday dates, each of the subjects counted in C_x^t and C_x^{t+1} will be on average ½ year at risk of dying. On the other hand, we will subtract/add the exact time that deceased or migrants have been/have not been exposed to risk as members of the population with age x during year *t*. In particular, from $\frac{1}{2}C_x^t$ we will deduct the time that deceased and migrants have not been exposed to risk, i.e. $\sum_{j=1}^{D_{x:U}^t}(1-l_{x,j}^t)+\sum_{j=1}^{E_{x:U}^t}(1-l_{x,j}^t)$, and we will add the unaccounted time of risk of the immigrants, i.e. $\sum_{j=1}^{I_{x:U}^t} i_{x,j}^t$. Similarly, we will add $\sum_{j=1}^{D_{x:L}^t} d_{x,j}^t+\sum_{j=1}^{E_{x:L}^t} e_{x,j}^t$ and take away $\sum_{j=1}^{I_{x:L}^t}(1-l_{x,j}^t)$ from $\frac{1}{2}C_x^{t+1}$. Thus, in this scenario, the mortality rate estimator, $\widetilde{m}_x$, would be expressed by equation (4). As in the previous case, the estimator for the probability of death will be: $\ddot{q}_x=\frac{\ddot{m}_x}{1+(1-\ddot{f}_x)\ddot{m}_x}$.

$$\ddot{m}_x=\frac{D_x^t}{\ddot{E}\dot{R}_{ABC}+\ddot{E}\dot{R}_{ACD}} \tag{4}$$

where $\ddot{E}\dot{R}_{ABC}=\frac{1}{2}C_x^t-\sum_{j=1}^{D_{x:U}^t}(1-l_{x,j}^t)-\sum_{j=1}^{E_{x:U}^t}(1-l_{x,j}^t)+\sum_{j=1}^{I_{x:U}^t} i_{x,j}^t$ and $\ddot{E}\dot{R}_{ACD}=\frac{1}{2}C_x^{t+1}+\sum_{j=1}^{D_{x:L}^t} d_{x,j}^t+\sum_{j=1}^{E_{x:L}^t} e_{x,j}^t-\sum_{j=1}^{I_{x:L}^t}(1-l_{x,j}^t)$.

2.5. Open population with no hypotheses about the distribution of deaths, migrants and births (OP_NUD_NUM_NUB)

The scenario that considers a greater level of detail in the data is that in which we have microdata on (i) the dates of death, (ii) the dates of immigration and emigration, and (iii) birth dates of the population. The provision of the exact date of birth of each member of the population allows us to measure more precisely the time exposed to risk that each individual carries.

In the previous scenarios, we have worked with the hypothesis that the birth dates of the population are distributed evenly throughout all the days of the year. Several studies, however, have questioned the validity of such a hypothesis, showing that the phenomenon is heavily influenced by environmental and cultural issues (e.g.,

Lam and Miron, 1994; Bobak and Gjonca, 2001). Hence, the estimators used so far may be inefficient (and as can be seen in the annex inconsistent).

The new estimator we propose incorporates, in addition to the exact time exposed to risk of deaths, immigration and emigration (analyzed in the previous points), the exact time exposed to risk of each individual who remains alive with age x during year t as a member of the target population. For example, a person born on January 2 of year $t-x-1$ (i.e., accounted in $C_{x,2}^{t}$) who does not die or migrate during year t will be exposed to the risk of dying, in a non-leap year, for $\frac{(2-0.5)}{365}$ years. Similarly, a person born, for example, on February 3 of year t −x (i.e., accounted in $C_{x,34}^{t+1}$) will be exposed to $\frac{(365-34+0.5)}{365}$ years, assuming the person has remained a member of the risk population throughout their lifetime with age x during year t. The 0.5 of the previous expressions arises from assuming that throughout each day the births are evenly distributed. That is to say, on average, those born on any given day do so at noon.

By adding the above quantities for each subject of $C_{x,d}^{t}$ and $C_{x,d}^{t+1}$ we will have the total time exposed to risk of the persons accounted for in C_{x}^{t} and C_{x}^{t+1} as if none of their lifelines had experienced any death or migration events within the ABCD square. However, in both triangles, ABC and ACD, in which ABCD is divided, events can occur that affect the different lifelines, making it necessary to make some adjustments to the aggregation. On the one hand, those members of C_{x}^{t} who die or emigrate within the triangle ABC have been counted as if they were to remain at risk all the time, when this is not the case. So, for each one of them it is necessary to discount the time that they have not remained members of the risk population; that is, $1-l_{x,j}^{t}$ years. Conversely, those individuals who immigrated to ABC have not been counted in C_{x}^{t}, so for them it is necessary to add the time each of them is at risk of dying: $i_{x,j}^{t}$. On the other hand, in the ACD triangle we find people who are not counted in C_{x}^{t+1} because they died or emigrated before January 1 of $t+1$, but as they had been exposed to the risk of dying as members of the target population, we must add the times, $d_{x,j}^{t}$ and $e_{x,j}^{t}$, that each of them was exposed to risk. Finally, it is necessary to subtract for each of the immigrants that have been included in the segment C_{x}^{t+1} the time that they have not been exposed

to risk of dying as members of the study group, i.e., $l_{x,j}^t$ years. Hence, in this scenario, the death rate is estimated according to the following equation (5):

$$\breve{m}_x = \frac{D_x^t}{\breve{ER}_{ABC} + \breve{ER}_{ACD}} \tag{5}$$

Where $\breve{ER}_{ABC} = \sum_{d=1}^{T} \frac{(d-0.5)}{T} C_{x,d}^t - \sum_{j=1}^{D_{x:U}^t} (1 - l_{x,j}^t) - \sum_{j=1}^{E_{x:U}^t} (1 l_{x,j}^t) + \sum_{j=1}^{I_{x:U}^t} i_{x,j}^t$, $\breve{ER}_{ACD} = \sum_{d=1}^{T} \frac{(T-d+0.5)}{T} C_{x,d}^{t+1} + \sum_{j=1}^{D_{x:L}^t} d_{x,j}^t + \sum_{j=1}^{E_{x:L}^t} e_{x,j}^t - \sum_{j=1}^{I_{x:L}^t} l_{x,j}^t$, d is the birthday and $\mathrm{T} = 365$, except in leap years when it is equal to $\mathrm{T} = 366$. In line with the two previous scenarios, the probability of death is obtained with: $\breve{q}_x = \frac{\breve{m}_x}{1+(1-\breve{f}_x)\breve{m}_x}$.

3. Data and statistical tests

The knowledge of the exact moment of the events of death, immigration and emigration and birth dates has allowed the development, within the family of the period-based estimators, of new estimators for the mortality rate, m_x. The estimators (1)–(4) of the previous section rest on a series of hypotheses that would be interesting to test. In the fourth section, we evaluate the suitability of the hypotheses on an actual database. In this section, we introduce the database and we detail the statistical hypothesis tests used to evaluate them.

3.1. Data and software

The microdata used for the present research comes from the population of the Comunitat Valenciana (Spain), with more than 5 million inhabitants. In particular, microdata by sex for the years 2010 to 2013 corresponding to dates of death, dates of immigration and emigration and dates of birthdays have been used. The microdata of deceased (day, month and year of death and birth) and migratory flows (day, month and year of migration and birth) have been provided by the Spanish National Official Institute (INE). The data corresponding to population registers (day, month and year of birth) have been provided by the Valencian Institute of Statistics (IVE).

The analysis of population trends is critically dependent on the quality of the data (Cairns *et al.*, 2016), and mainly on migration data. Although Spanish migration data are generally reliable and consistent, as Lledó *et al.* (2016) have proven, migration statistics have a reputation of being problematic and inconsistent (Wísniowski *et al.*, 2016), and

this shows through in some limitations in the Spanish data. The limitations of Spanish migration data come from the practice followed by INE agents of setting the first day of January as the birth date for those immigrants that do not know the exact date of her/his birth. This provokes an artificial peak in the data. Hence, before using immigrants' microdata, a number of first-of-January-born immigrants, equivalent to the artificial excess (measured as the difference over the corresponding daily mean of immigrants of the year), were randomly selected and randomly assigned a date of birth. The datasets with the amended immigrant data were the ones used to run the tests and estimate the life tables. All the computations have been performed using the statistical software R, version 3.3.0 (R Core Team, 2016).

3.2. Statistical tests for death and migration hypotheses

This section introduces the statistical tests used to evaluate the hypotheses of uniform distribution of deaths and migratory flows. These are made up of three types: spatial, functional and parametric. The closed demographic system hypothesis has not been tested. In contrast to *cohort-based estimators*, the equivalence between equations (1) and (2) makes in this case its contrast unnecessary.

From a geometric point of view, the knowledge of the dates of each demographic event enables accurate placement of deaths, migrations and emigrations (crosses, circles and squares) in each geometric figure (squares and triangles) of the Lexis diagram and to envisage the set of points where crosses, circles and squares occur as realizations of bivariate point pattern processes (year, age) in the Cartesian plane. From a spatial point of view, the hypothesis of uniformity in squares and triangles is equivalent to the hypothesis of complete spatial randomness (henceforth CSR). CSR hypothesis have been tested using some of the spatial tests available in version 1.47-0 of the R package *spatstat* (Baddeley and Turner, 2005). In particular, we have used the CLF test (Cressie, 1991; Loosmore and Ford, 2006), the Maximum Absolute Deviation (MAD) test (Ripley, 1977, 1981) and a spatial version of the well-known chi-squared goodness-of-fit (XS) test. For the XS test, we have divided triangles and squares in, respectively, eight and sixteen equal parts and measured the well-known Pearson $\mathcal{X}^2$ statistic distance between the number of observed and expected points.

In addition to the spatial tests, we have also used two functional tests. The objective is to compare the compatibility between the theoretical distribution of times exposed (non-exposed) to risk derived from the hypotheses of uniformity and the empirical distributions of times exposed (non-exposed) to risk obtained in each triangle for each type of demographic event. Under the hypothesis of uniformity, the density function of the random variable τ that measures the time exposed (non-exposed) to risk of deceased, immigrants and emigrants in each triangle is $f(\tau) = 2 - 2\tau$, for $0 \leq \tau < 1$ (Lledó *et al.*, 2016). The well-know one-sample Kolmogorov-Smirnov (KS) test (see, e.g., Conover, 1971) programmed in the R function *ks.test* and the geometric (G) test available in the R library *GoFKernel* (Pavia, 2015) have been the two functional tests used.

The completion of the different spatial and functional contrasts allows us to evaluate whether within each square and triangle the different demographic events are evenly distributed. It could be, however, that even if such hypotheses are not generally acceptable, their concreteness in equations (1)–(4) might be. Hence, in addition to the previous tests, we have implemented an additional battery of parametric tests to evaluate the suitability of the formulas. On the one hand, we evaluate whether the average time of exposure (non-exposure) to risk of people dying or migrating is $\frac{1}{3}$ in both triangles. On the other hand, we study whether the number of deceased, emigrants and immigrants is equal in the two triangles. One-sample two-side mean t-tests and binomial tests (with $p = 0.5$) are used to check the above hypotheses. Finally, given that if the total time exposed/not-exposed to risk of deaths, emigrants and immigrants is equal in lower and upper triangles (i.e., if $D_{x:U}^t - \sum_{i=1}^{D_{x:U}^t} l_{x,i}^t = \sum_{i=1}^{D_{x:L}^t} d_{x,i}^t$, $E_{x:U}^t - \sum_{i=1}^{E_{x:U}^t} l_{x,i}^t = \sum_{i=1}^{E_{x:L}^t} e_{x,i}^t$ and $\sum_{i=1}^{I_{x:U}^t} i_{x,i}^t = I_{x:L}^t - \sum_{i=1}^{I_{x:L}^t} l_{x,i}^t$) equations (3) and (4) collapse to equation (1), we have also tested using bootstrap (Efron, 1993) the suitability of the above equalities.

3.3. Statistical tests for the hypothesis of uniform distribution of births

The last hypothesis analyzed and evaluated in this paper is the hypothesis of uniform distribution of dates of birthdays (births) in each group C_x^t and C_x^{t+1}. This hypothesis is widely accepted by government agencies when calculating mortality rates (e.g., ABS, 2016; Arias, 2015; INE, 2016; ONS 2010, 2012). Despite that, some studies warn "this

assumption is violated most severely in situations where there are rapid changes in the size of successive cohort, owing to fluctuations in the birth series many years before" (Wilmoth *et al.*, 2007, p. 79).

To assess this hypothesis, we evaluate: (i) whether the monthly proportions of birth dates are distributed according to a multinomial distribution with size C_x^t (or C_x^{t+1}) and probability vector $\frac{(31,F,31,30,31,30,31,31,30,31,30,31)}{T}$, where $F = 28$ and $T = 365$ (except when $x - t - 1$ ($x - t$) is a leap year when F and T are, respectively, 29 and 366); and (ii) whether the distribution of birth dates in each year follows a continuous uniform distribution. To evaluate (i) we use the $\mathcal{X}^2$ goodness-of-fit test, while to evaluate (ii) we use the KS and G tests introduced in the previous subsection, after transforming the birthdays into a continuous variable. Specifically, each person born on day dth (i.e., included in $C_{x,d}^t$ or $C_{x,d}^{t+1}$) was randomly assigned a birth moment within day d, that is, in the interval $]d - 1, d[$.

The hypothesis of even distribution of birthdays, however, is not the only assumption that could lead to a mean ½ year risk exposure time for subjects counted in C_x^t and C_x^{t+1}. Distributions in the form of U, U-inverted or in general any symmetrical distribution would also lead to the same result. Therefore, as a parametric proxy we also test whether the proportion of birthdays in each half of the year is ½.

4. Results of the statistical tests

This section shows the results and conclusions reached after applying the statistical tests detailed in the previous section to the data sets of demographic events of the Comunitat Valenciana (Spain) corresponding to the years 2010 to 2013. Given the enormous amount of tests made (more than 55,000), and in order not to overload the presentation of the results with an excessive number of figures, this work comes accompanied by supplementary graphical material.

4.1. Testing uniformity in the distribution of deaths

The hypothesis of uniform distribution of deceased is widely used in the construction of life tables. To evaluate it, we obtained the p-values of the XS, MAD and CLF spatial tests and of the functional tests KS and G for each period between 2010 and 2013 and each age between 0 and 108 years in each of the triangles and squares of the Lexis scheme.

The results of the tests for the upper triangles (4,360 p-values: 109 years, 5 tests, 2 genres and 4 years) are shown in Figure 2. Those p-values surpassing the significant levels of 0.01, 0.05 and 0.10 are coloured in a gray scale. The values of the uniformity tests performed for lower and square triangles —which draw conclusions similar to those of the upper triangles— are available in Figures 1-S and 2-S in the supplementary material.

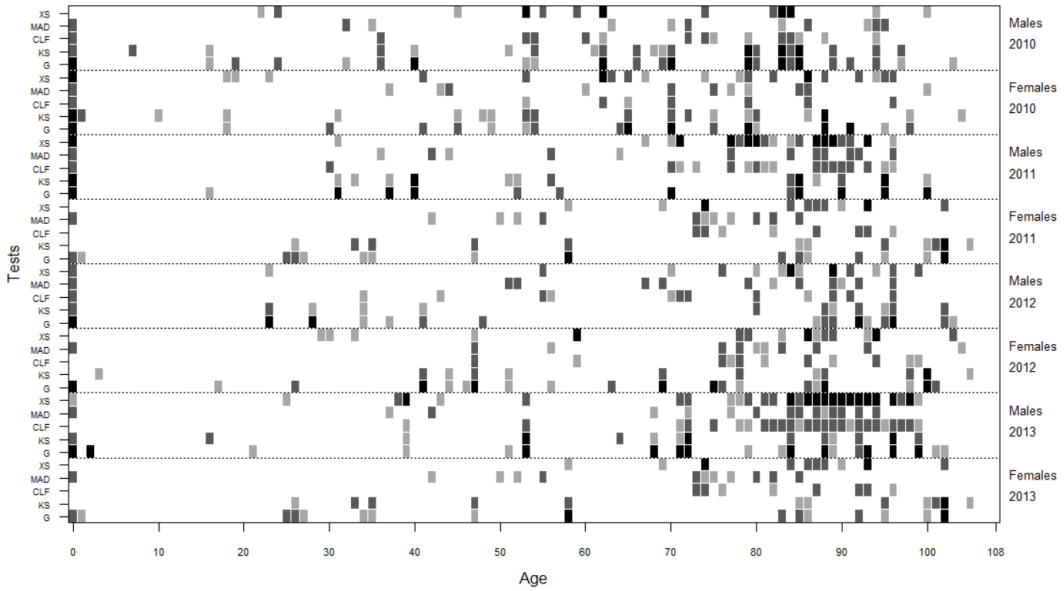

Figure 2. Uniform hypothesis tests by gender and age in Comunitat Valenciana population for people dying in 2010-2013 Lexis upper triangles. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.10 significant levels. MAD denotes the MAD test, CLF the Cressie–Loosmore–Ford test, XS the spatial Chi-squared goodness-of fit test, KS the Kolmogorov–Smirnov test and G the Geometric test. The first three tests are spatial tests and check CSR as null hypothesis after a representation of events in the Lexis space. KS and G tests are functional nonparametric tests and check whether the empirical distributions of non-exposed-to-risk times are compatible with the assumed probability distribution.

The results of the different tests indicate that it is generally correct to assume the hypothesis of uniformity within the different triangles and squares for the young and adult population, where mortality is low. As age increases, however, the validity of this hypothesis could be questioned. From the age of 70 years, the number of rejections of the hypothesis visibly grows. The age of 0 years deserves special attention, with a widespread rejection, as a consequence of the well-documented rise in the number of deceased in the first days/weeks of life.

In equations (1) and (2), the hypothesis of the uniformity of the deceased resulted in that, on average, each deceased person would be (would not be) $\frac{1}{3}$ year at

risk of dying with age x during year t and that the number of deceased would be equal in the lower and upper triangles. Figure 3 shows for our database the results of such tests, as well as the result of the combination of both hypotheses, which states that the times of exposure/non-exposure during year t by the set of deaths with age x in both triangles are equal.

Figure 3 shows that it is generally acceptable, with less intensity from the age of 70, that the time exposed (non-exposed) to risk of the deceased is $\frac{1}{3}$ in both triangles (see also Figures 9-S and 10-S in the supplementary material). The only generalized rejection occurs for the age of 0 years, where the average time lived by the deceased, f_0^t, tends to differ from the expected value ½. For instance, for males in 2013, we have: $f_0^{2013} = 0.199$. A similar result is obtained when analyzing the corresponding p-values of the hypothesis of an equal number of deaths in both triangles. However, when the joint compatibility of both hypotheses is studied, that is, when we study if the total times exposed to risk of the deceased in the lower triangles and that non-exposed of the deceased in the upper triangles are the same, the result is a resounding rejection.

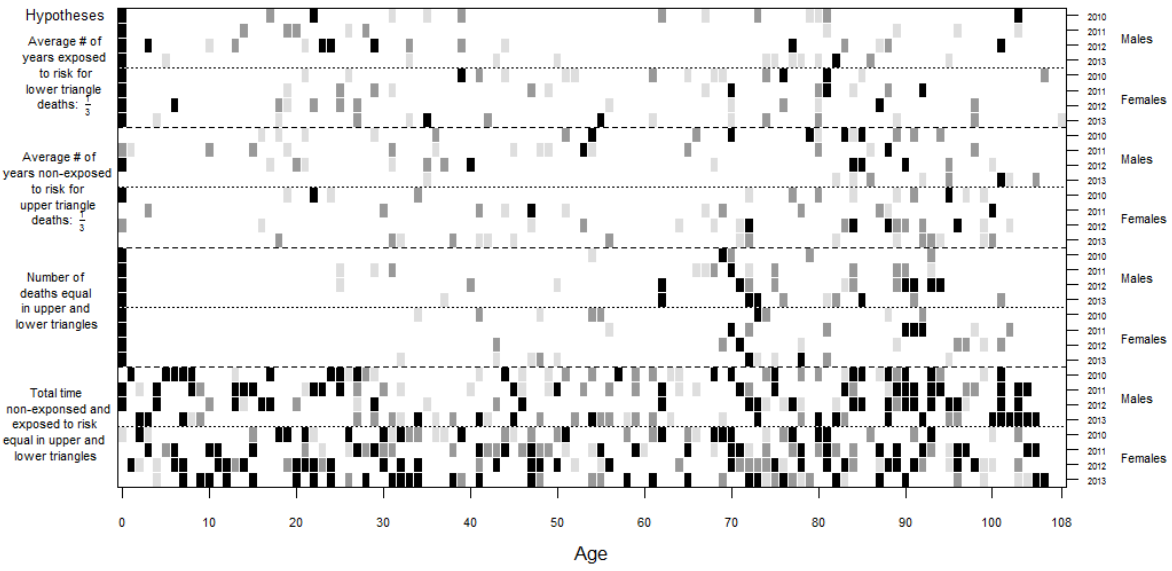

Figure 3. Parametric hypothesis tests, by gender and age for 2010–2013 Comunitat Valenciana population, corresponding to the concreteness in equations (1) and (2) of the hypotheses of uniform distribution of deaths. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.10 significant levels. Two-side mean t-student tests used to test the hypotheses of the first two blocks, two-side binomial tests with $p = 0.5$ to assess the hypotheses of the third block and a test based on the bootstrap empirical distribution of the differences to gauge the hypotheses of the last block.

4.2. Testing uniformity in the distributions of migration flows

Spain, and in particular the Comunitat Valenciana, has always been characterized by a high frequency in its migratory flows (Massey *et al.*, 1993; Cabrer and Pavía, 2003), mainly motivated by economic factors (Esipova *et al.*, 2011). Since the 1970s, Spain (and the Comunitat Valenciana) has been characterized as a country with a significant positive migratory flow. However, as a consequence of the economic crisis that began in 2007, the intensity of the flows has subsided and the balance has been reversed, becoming negative as of 2010, mainly as a result of an exodus of the younger population (see Figure 44-S in the supplementary material). In this section, we evaluate: (i) whether for each age x and period t the migratory flows are evenly distributed within squares and triangles; (ii) whether the time exposed (non-exposed) to risk of immigrants and emigrants is $\frac{1}{3}$ in each triangle; (iii) whether the numbers of immigrants (emigrants) in both triangles are equal; and (iv) whether the total time exposed (non-exposed) to risk of immigrants (emigrants) is equal in both triangles.

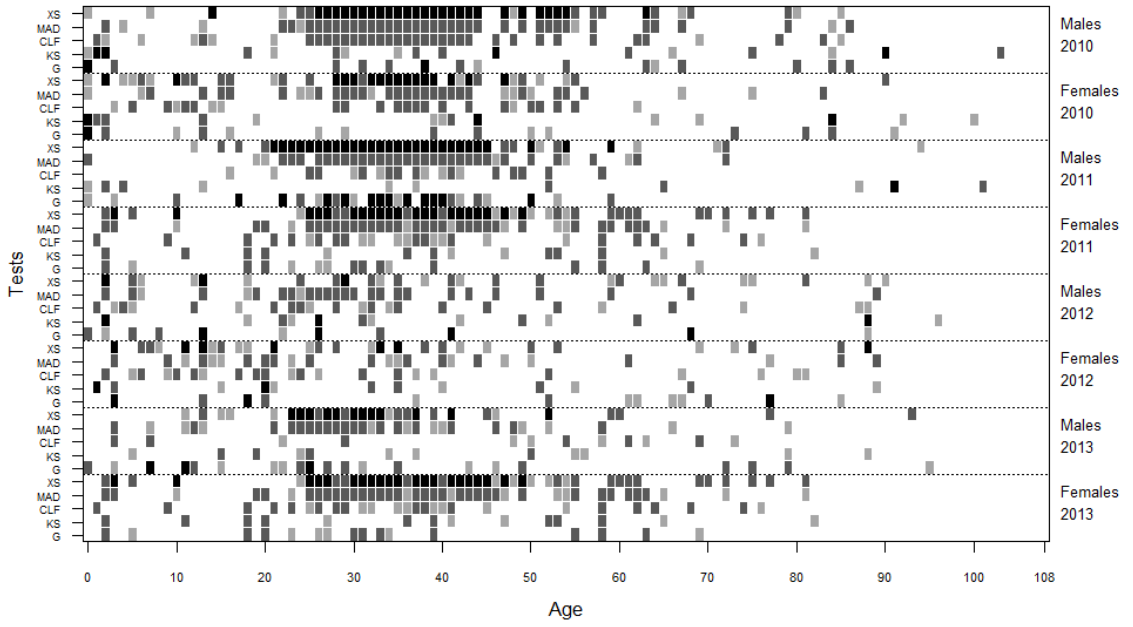

Figure 4. Uniform hypothesis tests by gender and age in the Comunitat Valenciana population for emigrant events occurring in 2010-2013 Lexis lower triangles. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.10 significant levels. MAD denotes the MAD test, CLF the Cressie–Loosmore–Ford test, XS the spatial Chi-squared goodness-of fit test, KS the Kolmogorov–Smirnov test and G the Geometric test. The first three tests are spatial tests and check CSR as null hypothesis after a representation of events in the Lexis space. KS and G tests are functional nonparametric tests and check whether the empirical distributions of exposed-to-risk times are compatible with the assumed probability distribution.

As an example, in Figure 4 we show the results of the spatial and functional tests performed to test the hypothesis of uniform distribution of emigrants in lower triangles. The results for immigrants in squares and triangles (upper and lower) and for emigrants in squares and upper triangles are not offered here for reasons of space. They can be consulted in Figures 3-S to 7-S in the supplementary material. Overall, there are many ages and Lexis surfaces for which the results of the tests, especially spatial, point to a non-compliance with the hypothesis of homogeneity for migratory flows, although the intensity by age varies between types of flows and years. For example, rejections are more frequent in the age range up to 50 years in immigration flows and in the range 20-45 in those of emigration. By sex, the intensity of rejection of the null hypothesis is greater in men than in women.

The results obtained in Figure 5 (and Figures 8-S to 10-S in the supplementary material) generally show inconclusive results for the hypothesis that the time exposed (non-exposed) to risk of immigrants and emigrants is $\frac{1}{3}$ in both triangles, although the number of rejections is significant. In any case, as shown by the results of the migrant equality tests in both triangles and, above all, the tests of equality of the total times of exposure and non-exposure to risk among triangles, the impact of migratory flows is not compensated between the two components of risk exposure, so the estimator (4) would not collapse into estimator (3). The actual impact of these differences at each age, however, could be conditioned by the relative weight that net migratory flows had over the total corresponding population, which for the period analyzed were not particularly intense in the study area.

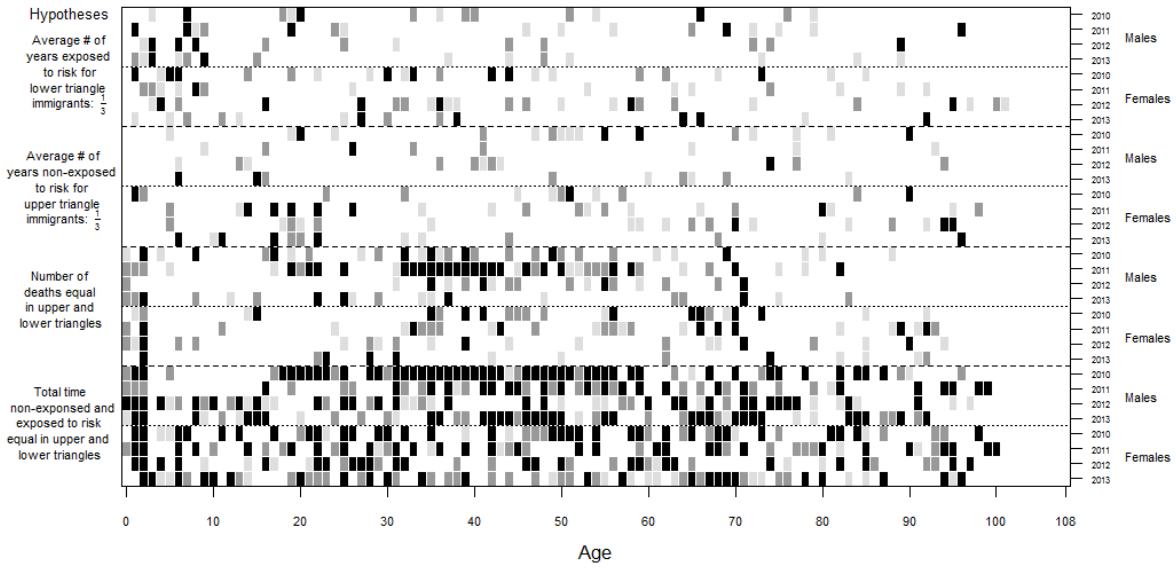

Figure 5. Parametric hypothesis tests, by gender and age for 2010–2013 Comunitat Valenciana population, corresponding to the concreteness in equations (1) and (2) of the hypotheses of uniform distribution of immigrants. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.10 significant levels. Two-side mean t-student tests are used to test the first two blocks of hypotheses, two-side binomial tests with $p = 0.5$ to assess the third block and a test based on the bootstrap empirical distribution to gauge the last block of hypotheses.

4.3. Testing uniformity in the distribution of birthdays

The hypothesis of uniform distribution of birthdays is probably the most implicit (and least studied) assumption of all of those that are commonly used in estimating the probabilities of death. Its accuracy is, however, dubious. Studies have shown that the annual distributions of births have evolved from environmentally regulated patterns to patterns governed by sociocultural factors (Quesada, 2006). In the Comunitat Valenciana, a pattern of births with a predominance of births in the first quarter of the year, for the 1940s and 1950s, has been changed to a scheme in which the highest concentration of births is recorded during the last quarter (see Figure 11-S in the supplementary material). A sudden or continuous increase in the number of births at specific times of the year can have a quantifiable impact on the life table.

Under the hypothesis of even distribution of birthdays, it is expected (i) that approximately the same number of births will occur in the first and second half of the year, (ii) that the proportion of the number of births recorded in each of the months is proportional to the number of days of the month, and (iii) that the number of births on each day of each year is approximately constant. Figure 6 shows the result of testing the above hypotheses. If in the previous analyzes some doubts existed about the plausibility

of the hypotheses, in this case the results obtained are overwhelming. The evidence clearly points to the fact that the hypothesis of uniform distribution of birthdays should be rejected.

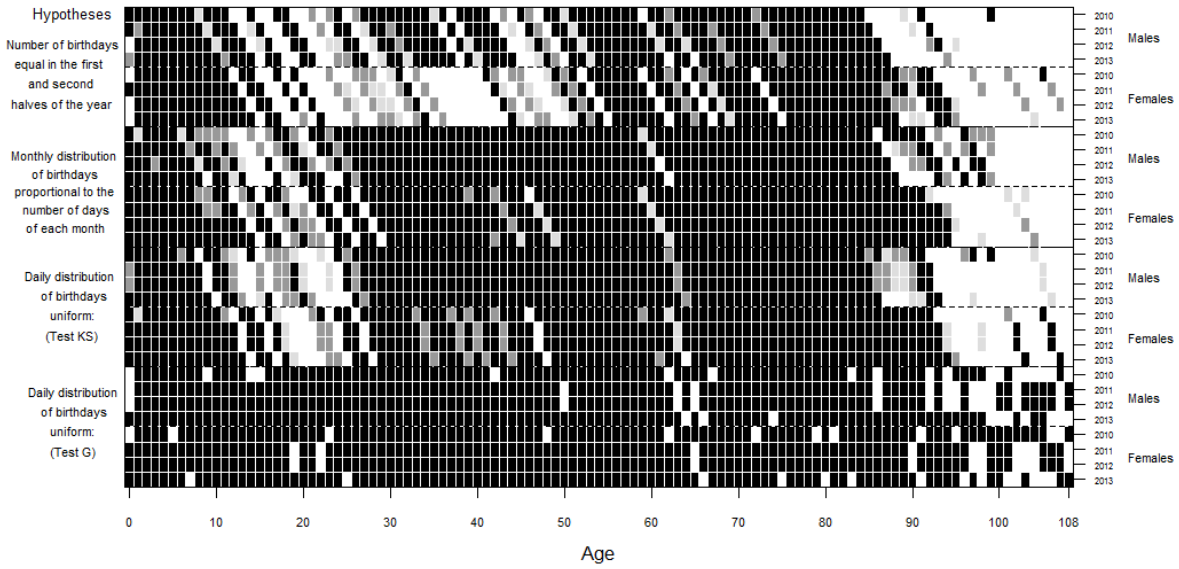

Figure 6. Statistical tests, by gender and age for 2010–2013 Comunitat Valenciana population, to assess the hypothesis of uniform distribution of birthdays. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.1 significant levels. Two-side binomial tests with $p = 0.5$ are used to test the first block of hypotheses, the second block of hypotheses is checked using $\mathcal{X}^2$ goodness-of-fit tests and non-parametric KS and G test are employed in, respectively, the last two blocks.

5. Comparing life tables

In the previous section, we have seen through a battery of statistical tests (spatial, functional and parametric) that the hypothesis of uniform distribution of birthdays is clearly inadequate and that there are also doubts about the global validity of the hypotheses of uniform distribution of deaths and migratory flows. It could be, however, that its effect on the life table might not be so significant. In this section, we study the impact of the hypotheses on the estimated death probabilities and on some insurance products (such as annuities and year-term life insurances). In particular, we compare the life tables attained using the different estimators and we study, through some examples, the differences that using each table entails in terms of monthly payments (for some temporal annuities) and in terms of premiums (for some year-term life insurances).

5.1. Impact of the hypotheses on death probabilities

The equations derived in the second section allow the estimation of the mortality rates and from these the probabilities of death. A total of 32 life tables (2 genres, 4 years and 4 different scenarios), each comprising of two sets of estimates, have been computed. This represents a total of 6,464 point estimates, after truncating for statistical stability reasons the estimated life tables in to 100 years. For the comparison of the different probabilities and rates a dissimilarity statistic has been used. Specifically, the Absolute Relative Discrepancy (ARD): $\frac{|\hat{q}_x - \tilde{q}_x|}{\tilde{q}_x}, x = 0,1 \dots 100$. Comparisons for both m_x and q_x have been computed.

Contrary to what happens with tables 2006-2008 obtained by Lledó *et al.* (2016) for Spain using *cohort-based estimators*, in our study the hypotheses of closed population and of uniform distributions of deaths and migration flows do not show any significant impact on the estimated tables (see Figures 13-S to 36-S in the supplementary material). The life tables obtained in the first four scenarios are almost equal. It seems that using populations corresponding to two different periods has a compensating effect on the migration flows which dilutes any potential impact. This contrasts with the conclusions reached when *cohort-based estimators* are used, where for example the comparisons between closed and open population scenarios show discrepancies surpassing even 4% (Lledó *et al.*, 2016).

Non-compliance with the hypothesis of uniform distribution of birthdays does, however, appear to have a measurable effect on the probability of death. For example, as can be seen in Figure 7, where the discrepancies are compared for the tables of 2013 between the OP_NUD_NUM_UB (open population with no uniform distribution for deaths and migrants and hypothesis of uniform distribution of birth dates) and OP_NUD_NUM_NUB (open population with no hypotheses), which differ only in the assumption of uniformity of births, applying such a hypothesis leads to differences of around 1.00%-1.50%. The greatest differences, between 3%-5%, are observed for the ages of 73 and 74, which correspond to the cohort of births a few months after the end of the Spanish Civil War. This extraordinary historical event provoked a higher concentration of births in late 1939 and early 1940. This same phenomenon and effect on the mortality table is predictably present in the tables of those countries that

participated in World War I and II. The other peak observed in Figure 7 corresponds to those born in 1918, the year of the great Spanish flu that killed millions of people around the world and that in Spain resulted in a severely reduced level of births in the months of July and August.

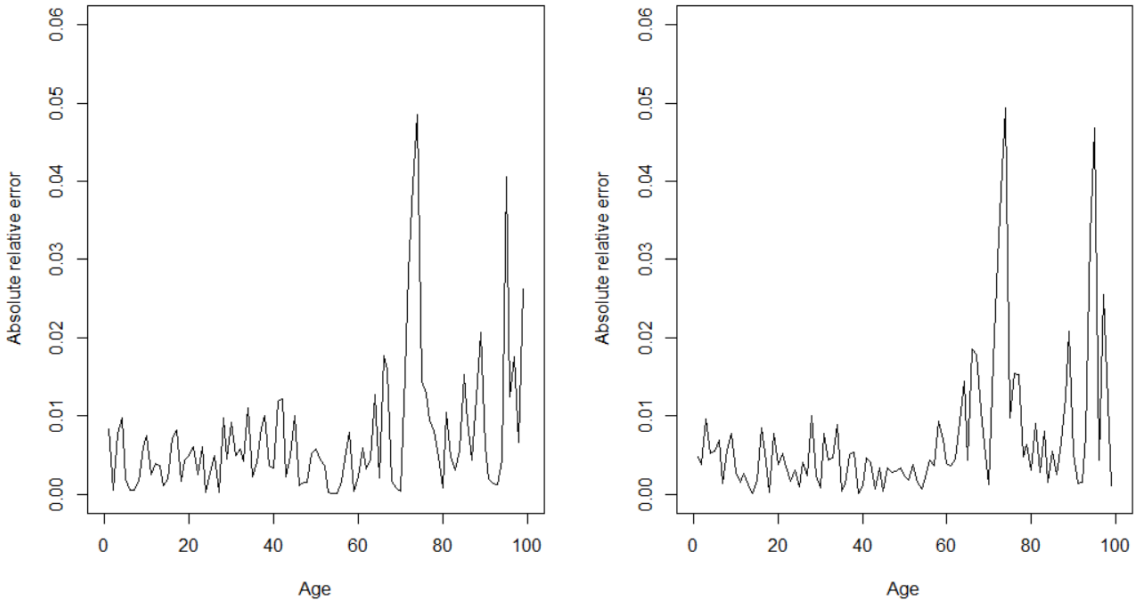

Figure 7. Absolute relative discrepancies between the probabilities of death for 2013 of the OP_NUD_NUM_UB (open population with no hypotheses of distribution of deaths and migrants and uniform distribution of births) and the OP_NUD_NUM_NUB scenarios (open population with no hypotheses about distribution of deaths, migrants and of births) scenarios. Left panel: Discrepancies of men life table. Right panel: Discrepancies of women life table.

5.2. Impact of the hypotheses on insurance products

The life table is used in different areas of the demographic and actuarial field. For example, in life insurance companies, it is used to calculate the premium payable by an insurer and, in the public pension systems, it is used for calculating the pension to be received. To assess the impact of different assumptions on insurance products, by way of example, we calculate for some combinations of age and time some annuities and premiums. The insurance product with the highest commercialization are annuities. Table 3 shows the differences between the actuarial values of a monthly annuity, ${}_t\ddot{a}_x^g$, for a person of age x payable during the next t years corresponding to the generation born in year g, for the tables that are derived from the OP_NUD_NUM_NUB and OP_NUD_NUM_NUB scenarios. This allows the evaluation of the impact of just the hypothesis of uniform distribution of births. The calculations are differentiated by gender for the sake of comparison. Currently, and because of the Test-Achats case,

gender cannot be used in the EU to discriminate premiums and benefits under insurance contracts. The selected ages were 35 and 55 years and the temporalities 30 and 10 years, respectively. We selected two completely contiguous generations (1940 and 1941) with the aim of also assessing the differences between generations.

Table 3 shows the differences in actual value of the different annuities considered. In addition to the differences between scenarios, differences between generations could also be calculated. Although the differences between the different annuities are not particularly high, as shown in Table 3, the hypothesis of uniform distribution of births is not innocuous. Its effect, however, is more evident for term life insurances.

Table 3. Examples of annuities for a total paying sum insured of 100,000 €.

	Men				Women			
Estimator	${}_{30}\ddot{a}_{35}^{1941}$	${}_{30}\ddot{a}_{35}^{1940}$	${}_{10}\ddot{a}_{55}^{1941}$	${}_{10}\ddot{a}_{55}^{1940}$	${}_{30}\ddot{a}_{35}^{1941}$	${}_{30}\ddot{a}_{35}^{1940}$	${}_{10}\ddot{a}_{55}^{1941}$	${}_{10}\ddot{a}_{55}^{1940}$
OP_NUD_NUM_UB	293.14 €	294.67 €	871.31 €	874.79 €	284.80 €	285.73 €	848.39 €	850.55 €
OP_NUD_NUM_NUB	293.74 €	293.92 €	872.86 €	872.81 €	285.07 €	285.39 €	849.01 €	849.76 €
Difference	-0.20%	0.26%	-0.18%	0.23%	-0.10%	0.12%	-0.07%	0.09%

The discount rate and expenses are assumed null. OP_NUD_NUM_UB: Open population with no hypotheses about distribution of deaths and migrants and uniform distribution of births. OP_NUD_NUM_NUB: open demographic system with no hypotheses about distribution of deaths, migrants and of births.

Table 4 shows the differences of premiums to be paid for a person of age x (with x = 35, 45, 55, 60 and 65) for a year-term life insurance of € 100,000. Again, the calculations are differentiated by gender for the sake of comparison. In view of the results and for the ages considered, the hypothesis of uniform distribution of births tends to reduce the premiums of this type of insurance. For example, the price of contracting a year-term life insurance for a 60-year-old woman would be 1.95% cheaper using the table obtained under the OP_NUD_NUM_NUB scenario than using the equivalent table with a uniform distribution of dates of birth.

Table 4. Premium to buy a year-term life insurance of 100,000 €.

	Men					Women				
Estimator	${}_1A_{35}$	${}_1A_{45}$	${}_1A_{55}$	${}_1A_{60}$	${}_1A_{65}$	${}_1A_{35}$	${}_1A_{45}$	${}_1A_{55}$	${}_1A_{60}$	${}_1A_{65}$
OP_NUD_NUM_UB	71.72 €	277.51 €	592.79 €	958.14 €	1,443.72 €	50.61 €	108.16 €	221.29 €	371.20 €	469.22 €
OP_NUD_NUM _NUB	72.67 €	278.29 €	596.96 €	976.48 €	1,464.14 €	50.80 €	107.80 €	222.60 €	378.60 €	476.90 €
Difference	-1.30%	-0.28%	-0.70%	-1.88%	-1.39%	-0.37%	0.34%	-0.59%	-1.95%	-1.61%

The expenses are assumed null. OP_NUD_NUM_UB: Open population with no hypotheses about distribution of deaths and migrants and uniform distribution of births. OP_NUD_NUM_NUB: open demographic system with no hypotheses about distribution of deaths, migrants and of births.

Also, although they are not shown in the tables, these differences increase even more at older ages. Specifically, for the age of 68, the differences are up to 4.68% for men and 5.05% for women, and even more if we consider people born in 1939 or 1940. Given the volume of this market, a change in the price of the premiums around the level of percentages this research shows could be a matter of great importance that could impact on each Company differently depending on its portfolio profile.

6. Conclusions

The mortality table summarizes the vital behavior of individuals who make up a given population (general or insured) over a period of time and in a particular region. In general populations, the table is constructed by official bodies (such as, the Office for National Statistics in the United Kingdom or the National Institute of Statistics in Spain). In insured populations, this work falls on insurance companies.

Currently, almost all statistical agencies construct life tables using estimators from the period-based family, after assuming some hypotheses about the behaviour of the demographic events. These estimators, however, show some theoretical inconsistencies. On the one hand, they (almost) always include deaths in the numerator of m_x that have not even been considered in the population exposed-to-risk when the average population at risk of dying is used as denominator of m_x. On the other hand, as we show in this paper (see the annex), when the total 'person-years' at risk is used as denominator of m_x, different mathematical expressions are possible for m_x depending on the reasoning followed to construct the estimator. In this paper, we propose an assumption-free estimator that exploits all the big microdata available and is free of the theoretical inconsistencies that is shown by the estimators currently in use by most official statistical agencies.

In addition to our theoretical contribution, we test with a real database the suitability of the hypotheses of uniform distribution of deaths, migrants and birthdays that are usually characteristic of period-based estimators and analyze their impact on the life table and on some insurance products. In light of the results, we conclude that assuming uniform distribution of birth dates is a hypothesis not supported by data and which has a measurable impact on the life table, with potentially deleterious effects on public pension systems and life insurances. Therefore, although the impact may not

exceed differences of 1.00-1.50%, except in exceptional circumstances, we see no reason for not adopting more efficient approaches, such as those proposed in this investigation. The technology and data are available. Indeed, we advocate that their use become common practice.

Finally, it should be noted that, although this work focuses on analyzing the impact of the three hypotheses (H1)–(H3) mentioned in the introduction, the construction of mortality tables usually includes a fourth implicit hypothesis that has not been considered and whose analysis is qualitatively much more complex. Specifically, in both period-based estimators and cohort-based estimators, it is implicitly assumed that immigrants (emigrants) acquire (possess) have the same risk of death as the resident population of the same age and sex and, therefore, there is no selection effect in the migration decision, with the risk of death at each age independent of the migrant's personal background. Although this hypothesis does not seem reasonable, its evaluation and analysis is extremely complex. For this, it would be necessary to have longitudinal data of the migrants, information that, at least in the case of Spain, is not publicly available. We encourage researchers who have access to this type of data to study their suitability and possible impact on estimations.

ANNEX: Alternative estimators for m_x

Equations (1)–(5) show the estimators obtained for m_x from stock populations as of January 1 of years *t* and *t+1*. In this annex, we show the equations (1)–(4) that would have been obtained if in each of the corresponding scenarios other segments of ABCD had been taken as starting elements. Equation (5) is invariant under all reasoning. In particular, we study the effect of two reasonable alternatives.

Alternative I. In line with Wilmoth *et al.* (2007), the AD and BC segments are considered as starting elements, that is, this alternative starts from the number of people who reach the exact age x and the exact age *x+1*, respectively, throughout year *t*.

Alternative II. This alternative starts with the total people who are still alive after passing throughout the area of exposure to risk, i.e. the BC and CD segments.

According to our notation the respective values of the segments AD, BC and CD are: $\mathrm{AD} = C_x^{t+1} + D_{x:L}^t + E_{x:L}^t - I_{x:L}^t$, $\mathrm{BC} = C_x^t - D_{x:U}^t - E_{x:U}^t + I_{x:U}^t$ and $\mathrm{CD} = C_x^{t+1}$.

Under Alternatives I and II and hypotheses of even distribution of birthdays, the number of 'person-years' at risk in the upper triangles would be: (i) ½ for each person counted in BC; (ii) plus the time exposed to risk of deceased and emigrants in ABC, since they are not accounted for in BC and yet have been exposed to risk; (iii) less the time that the immigrants in ABC have not been exposed to risk with age x during the year *t*, since they are accounted for in BC but have not been exposed to risk at all times.

On the other hand, in Alternative I, the number of 'person-years' at risk in the lower triangles will be: (i) ½ for each person accounted for in AD; (Ii) less the time of non-exposed to risk of the deceased or migrants in ACD; (Iii) plus the time the immigrants in ACD have been exposed to risk with age x during the year t. And in Alternative II, the number of 'person-years' at risk in the lower triangles will be: (i) ½ for each person accounted for in CD; (Ii) plus the time exposed to risk of deceased or migrants in ACD; (Iii) less the time the immigrants in ABC have not been exposed to risk with age x during the year *t.*

Scenario CP_UD_UB. Under closed population hypothesis and uniform distribution of the deceased for each age x and year t we have : (i) $D_{x:L}^t = D_{x:U}^t = \frac{1}{2}D_x^t$; (ii) $E_{x:L}^t = I_{x:L}^t = E_{x:U}^t = I_{x:U}^t = 0$; and, (iii) the average time exposed/non-exposed to risk of each deceased person in each triangle is $\frac{1}{3}$ of the year (Carstensen, 2007; Willmoth *et al.*, 2007; Pavía *et al.*, 2012; Lledó *et al.*, 2016). Hence, it follows that the estimators under Alternatives I and II are:

$$\widehat{m}_x^I = \frac{D_x^t}{\frac{1}{2}\left(C_x^t - \frac{1}{2}D_x^t\right) + \frac{1}{3}\cdot\frac{1}{2}D_x^t + \frac{1}{2}\left(C_x^{t+1} + \frac{1}{2}D_x^t\right) - \frac{1}{3}\cdot\frac{1}{2}D_x^t} = \frac{D_x^t}{\frac{1}{2}C_x^t + \frac{1}{2}C_x^{t+1}} \qquad (1.\mathrm{I})$$

$$\widehat{m}_x^{II} = \frac{D_x^t}{\frac{1}{2}\left(C_x^t - \frac{1}{2}D_x^t\right) + \frac{1}{3}\cdot\frac{1}{2}D_x^t + \frac{1}{2}C_x^{t+1} + \frac{1}{3}\cdot\frac{1}{2}D_x^t} = \frac{D_x^t}{\frac{1}{2}C_x^t + \frac{1}{2}C_x^{t+1} + \frac{1}{12}D_x^t} \qquad (1.\mathrm{II})$$

Note that (1.I) coincides with (1), although those exposed to risk in each triangle leading to the solution differ, and that (1.II) and (1) are different.

Scenario OP_UD_UM_UB. Under the open population hypothesis and uniform distribution of deaths and migrants for each age x and year *t* we have: (i) $D_{x:L}^t = D_{x:U}^t = \frac{1}{2}D_x^t$; (ii) $E_{x:L}^t = E_{x:U}^t = \frac{1}{2}E_x^t$; (iii) $I_{x:L}^t = I_{x:U}^t = \frac{1}{2}I_x^t$; and (iv) that the average time exposed/non-exposed to risk of each deceased/migrant in each triangle is $\frac{1}{3}$ of the year (Pavía *et al.*, 2012; Lledó *et al.*, 2016). Hence, it follows that the totals exposed to risk in the lower and upper triangles under Alternatives I and II are:

$$\overline{ER}_{ACD}^{I} = \frac{1}{2}\left(C_x^{t+1} + \frac{1}{2}D_x^t + \frac{1}{2}E_x^t - \frac{1}{2}I_x^t\right) - \frac{1}{6}D_x^t - \frac{1}{6}E_x^t + \frac{1}{6}I_x^t = \frac{1}{2}C_x^{t+1} + \frac{1}{12}D_x^t + \frac{1}{12}E_x^t - \frac{1}{12}I_x^t$$

$$\overline{ER}_{ACD}^{II} = \frac{1}{2}C_x^{t+1} - \frac{1}{6}D_x^t - \frac{1}{6}E_x^t + \frac{1}{6}I_x^t$$

$$\overline{ER}_{ABC}^{I} = \overline{ER}_{ABC}^{II} = \frac{1}{2}\left(C_x^t - \frac{1}{2}D_x^t - \frac{1}{2}E_x^t + \frac{1}{2}I_x^t\right) + \frac{1}{6}D_x^t + \frac{1}{6}E_x^t - \frac{1}{6}I_x^t = \frac{1}{2}C_x^t - \frac{1}{12}D_x^t - \frac{1}{12}E_x^t + \frac{1}{12}I_x^t$$

And consequently:

$$\bar{m}_x^I = \frac{D_x^t}{\frac{1}{2}C_x^t + \frac{1}{2}C_x^{t+1}} \qquad (2.\text{I})$$

$$\bar{m}_x^{II} = \frac{D_x^t}{\frac{1}{2}C_x^t + \frac{1}{2}C_x^{t+1} + \frac{1}{12}D_x^t + \frac{1}{12}E_x^t - \frac{1}{12}I_x^t} \qquad (2.\text{II})$$

Again (2.I) coincides with (2), arriving at the same solution by different paths, and (2.II) and (2) are again different.

Scenario CP_NUD_UB. Under closed population hypothesis and no assumptions about the uniform distribution of deceased, we have: (i) $E_{x:L}^t = I_{x:L}^t = E_{x:U}^t = I_{x:U}^t = 0$; (ii) that the total time exposed to risk of the deceased in the lower and upper triangles are, respectively, $\sum_{j=1}^{D_{x:L}^t} d_{x,j}^t$ and $\sum_{j=1}^{D_{x:U}^t} d_{x,j}^t$; and (iii) that the total time of non-exposed to risk of the deceased in the lower triangle is $\sum_{j=1}^{D_{x:L}^t} nd_{x,j}^t$, where $nd_{x,j}^t$ represents the time non-exposed to risk with age x during year *t* of the *j*th deceased in ACD, i.e., the temporal distance between their date of death and January 1 of year *t* + 1 (the length of the KD segment in the example of Figure 1-right). It follows that the estimators under Alternatives I and II are:

$$\widetilde{m}_x^I = \frac{D_x^t}{\frac{1}{2}(C_x^t - D_{x:U}^t) + \sum_{j=1}^{D_{x:U}^t} d_{x,j}^t + \frac{1}{2}(C_x^{t+1} + D_{x:L}^t) - \sum_{j=1}^{D_{x:L}^t} nd_{x,j}^t} \quad (3.\text{I})$$

$$\widetilde{m}_x^{II} = \frac{D_x^t}{\frac{1}{2}(C_x^t - D_{x:U}^t) + \sum_{j=1}^{D_{x:U}^t} d_{x,j}^t + \frac{1}{2}C_x^{t+1} + \sum_{j=1}^{D_{x:L}^t} d_{x,j}^t} \quad (3.\text{II})$$

In this scenario, both equations differ from equation (3). For them to match, it would have to be verified in scenario II: $\frac{1}{2}D_{x:U}^t - \sum_{j=1}^{D_{x:U}^t} d_{x,j}^t = D_{x:U}^t - \sum_{j=1}^{D_{x:U}^t} l_{x,j}^t$. And in scenario I, as well as verification in the previous equality, the following would also have to be verified: $\frac{1}{2}D_{x:L}^t - \sum_{j=1}^{D_{x:L}^t} nd_{x,j}^t = \sum_{j=1}^{D_{x:L}^t} d_{x,j}^t$. These are equalities that do not occur in practice.

Scenario OP_NUD_NUM_UB. Under the hypothesis of open population and no hypothesis on the distribution of deaths and migrants we have: (i) that the total times exposed to risk of the deceased, emigrants and immigrants in the lower and upper triangle are, respectively, $\sum_{j=1}^{D_{x:L}^t} d_{x,j}^t$, $\sum_{j=1}^{D_{x:U}^t} d_{x,j}^t$, $\sum_{j=1}^{E_{x:L}^t} e_{x,j}^t$, $\sum_{j=1}^{E_{x:U}^t} e_{x,j}^t$, $\sum_{j=1}^{I_{x:L}^t} i_{x,j}^t$ and $\sum_{j=1}^{I_{x:U}^t} i_{x,j}^t$; (ii) that the total times non-exposed to risk of the deceased and emigrants in the lower triangle are, respectively, $\sum_{j=1}^{D_{x:L}^t} nd_{x,j}^t$ and $\sum_{j=1}^{E_{x:L}^t} ne_{x,j}^t$, where $ne_{x,j}^t$ represents the time non-exposed to risk with age x during the year t of the j^{th} emigrant in ACD, i.e., the temporal distance between their date of emigration and January 1 of year t + 1 (the length of the segment QD in the example of Figure 1-right); and (iii) that the total time non-exposed to risk of immigrants in the upper triangle is $\sum_{j=1}^{I_{x:U}^t} ni_{x,j}^t$, where $ni_{x,j}^t$ represents the time non-exposed to risk with age x during year t of the j^{th} immigrant in ABC, i.e., the time distance between January 1 of year t and their date of immigration (the length of the BT segment in the example of Figure 1-right). Hence, it follows that the totals 'person-years' at risk in the lower and upper triangles under Alternatives I and II are:

$$\ddot{E}\dot{R}_{ACD}^I = \frac{1}{2}(C_x^{t+1} + D_{x:L}^t + E_{x:L}^t - I_{x:L}^t) - \sum_{j=1}^{D_{x:L}^t} nd_{x,j}^t - \sum_{j=1}^{E_{x:L}^t} ne_{x,j}^t + \sum_{j=1}^{I_{x:L}^t} i_{x,j}^t$$

$$\ddot{E}\dot{R}_{ACD}^{II} = \frac{1}{2}C_x^{t+1} + \sum_{j=1}^{D_{x:L}^t} d_{x,j}^t + \sum_{j=1}^{E_{x:L}^t} e_{x,j}^t - \sum_{j=1}^{I_{x:L}^t}(1 - l_{x,j}^t)$$

$$\ddot{E}\dot{R}_{ABC}^I = \ddot{E}\dot{R}_{ABC}^{II} = \frac{1}{2}(C_x^t - D_{x:U}^t - E_{x:U}^t + I_{x:U}^t) + \sum_{j=1}^{D_{x:U}^t} d_{x,j}^t + \sum_{j=1}^{E_{x:U}^t} e_{x,j}^t - \sum_{j=1}^{I_{x:U}^t} ni_{x,j}^t$$

And consequently:

$$\ddot{\dot{m}}_x = \frac{D_x^t}{\ddot{E}\ddot{R}_{ABC} + \ddot{E}\ddot{R}_{ACD}} \tag{4.I}$$

$$\ddot{\dot{m}}_x = \frac{D_x^t}{\ddot{E}\ddot{R}_{ABC} + \ddot{E}\ddot{R}_{ACD}} \tag{4.II}$$

For equations (4.I) and (4.II) to coincide with (4), the following equalities should be verified. For scenario II: $\frac{1}{2}D_{x:U}^t - \sum_{j=1}^{D_{x:U}^t} d_{x,j}^t = D_{x:U}^t - \sum_{j=1}^{D_{x:U}^t} l_{x,j}^t$, $\frac{1}{2}E_{x:U}^t - \sum_{j=1}^{E_{x:U}^t} e_{x,j}^t = E_{x:U}^t - \sum_{j=1}^{E_{x:U}^t} l_{x,j}^t$ and $\frac{1}{2}I_{x:U}^t - \sum_{j=1}^{I_{x:U}^t} ni_{x,j}^t = \sum_{j=1}^{I_{x:U}^t} i_{x,j}^t$. And for scenario I, as well as the previous equalities, also the equalities: $\frac{1}{2}D_{x:L}^t - \sum_{j=1}^{D_{x:L}^t} nd_{x,j}^t = \sum_{j=1}^{D_{x:L}^t} d_{x,j}^t$, $\frac{1}{2}E_{x:L}^t - \sum_{j=1}^{E_{x:L}^t} ne_{x,j}^t = \sum_{j=1}^{E_{x:L}^t} e_{x,j}^t$ and $\frac{1}{2}I_{x:L}^t - \sum_{j=1}^{I_{x:L}^t} i_{x,j}^t = I_{x:L}^t - \sum_{j=1}^{I_{x:L}^t} l_{x,j}^t$.

In practice, therefore, the hypothesis of uniform distribution of birthdays lead to the use of different estimators depending on the reasoning followed.

Supplementary material. Graphical Appendix (SA3)

Incorporating big microdata in life table construction: A hypothesis-free estimator

DEATH UNIFORM HYPOTHESIS TESTS

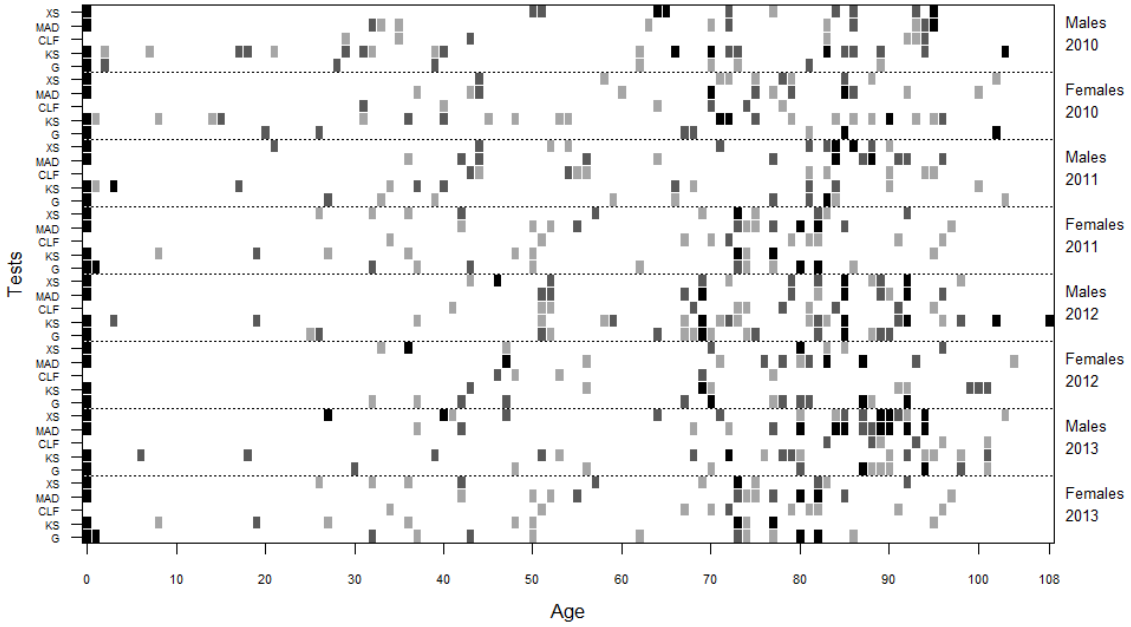

Figure 1-S. Uniform hypothesis tests by gender and age in Comunitat Valenciana population for people dying in 2010-2013 Lexis lower triangles. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.10 significant levels. MAD denotes the MAD test, CLF the Cressie–Loosmore–Ford test, XS the spatial Chi-squared goodness-of fit test, KS the Kolmogorov–Smirnov test and G the Geometric test. The three first tests are spatial tests and check CSR as null hypothesis after a representation of events in the Lexis space. KS and G tests are functional nonparametric tests and check whether the empirical distributions of exposed-to-risk times are compatible with the assumed probability distribution.

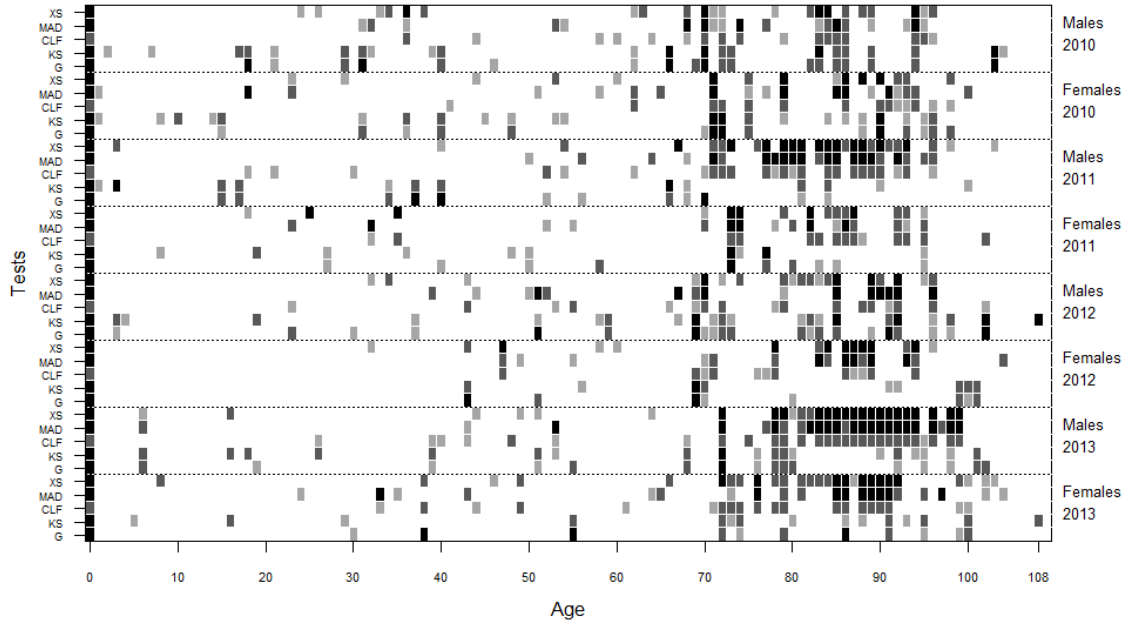

Figure 2-S. Uniform hypothesis tests by gender and age in Comunitat Valenciana (Spain) for people dying in 2010-2013 Lexis cells (squares). Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.10 significant levels. MAD denotes the MAD test, CLF the Cressie–Loosmore–Ford test, XS the spatial Chi-squared goodness-of fit test, KS the Kolmogorov–Smirnov test and G the Geometric test. The three first tests are spatial tests and check CSR as null hypothesis after

a representation of events in the Lexis space. KS and G tests are functional nonparametric tests and check whether the empirical distributions of exposed exposed-to-risk times are compatible with the assumed probability distribution; which for squares is equivalent to assume that the density function of the random τ variable that measures the time lived with completed age x for those dying in year t is $f(\tau) = 1$, for $0 \leq \tau < 1$.

IMMIGRANT UNIFORM HYPOTHESIS TESTS

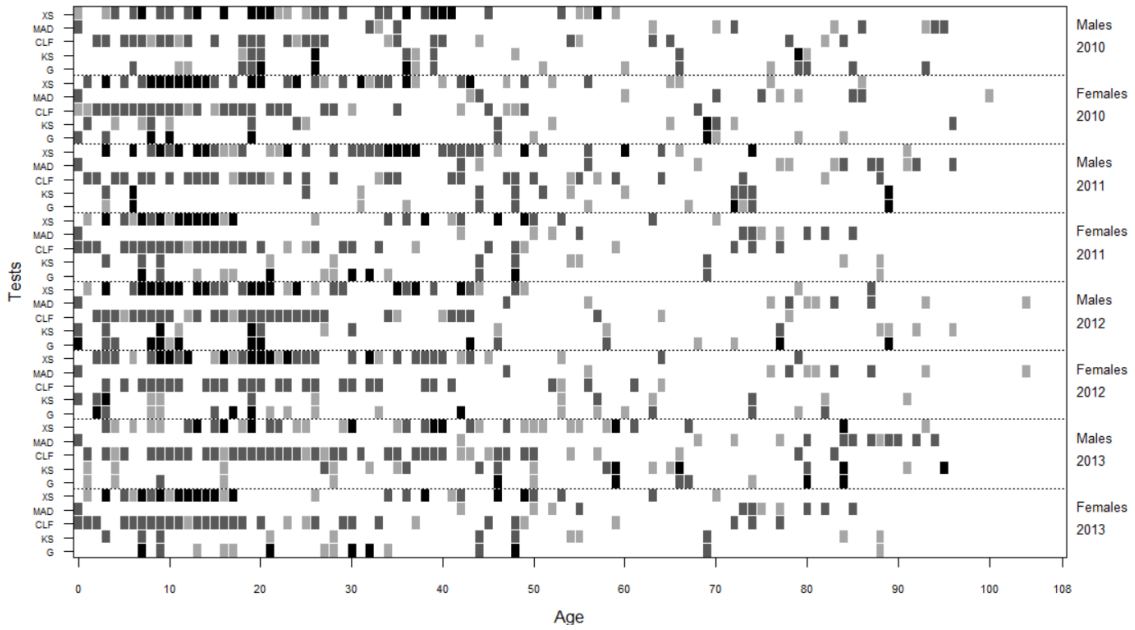

Figure 3-S. Uniform hypothesis tests by gender and age in Comunitat Valenciana population for immigrant events occurring in 2010-2013 Lexis low triangles. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.10 significant levels. MAD denotes the MAD test, CLF the Cressie–Loosmore–Ford test, XS the spatial Chi-squared goodness-of fit test, KS the Kolmogorov–Smirnov test and G the Geometric test. The three first tests are spatial tests and check CSR as null hypothesis after a representation of events in the Lexis space. KS and G tests are functional nonparametric tests and check whether the empirical distributions of non-exposed-to-risk times are compatible with the assumed probability distribution.

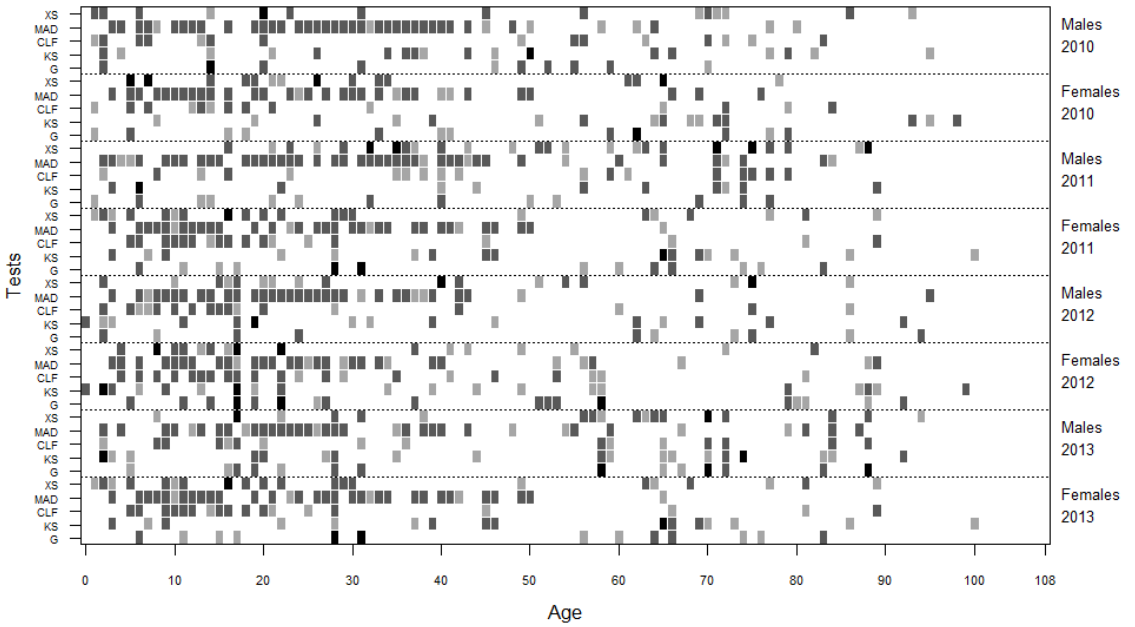

Figure 4-S. Uniform hypothesis tests by gender and age in Comunitat Valenciana population for immigrant events occurring in 2010-2013 Lexis upper triangles. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.10 significant levels. MAD denotes the MAD test, CLF the Cressie–Loosmore–Ford test, XS the spatial Chi-squared goodness-of fit test, KS the Kolmogorov–Smirnov test and G the Geometric test. The three first tests are spatial tests and check CSR

as null hypothesis after a representation of events in the Lexis space. KS and G tests are functional nonparametric tests and check whether the empirical distributions of exposed-to-risk times are compatible with the assumed probability distribution.

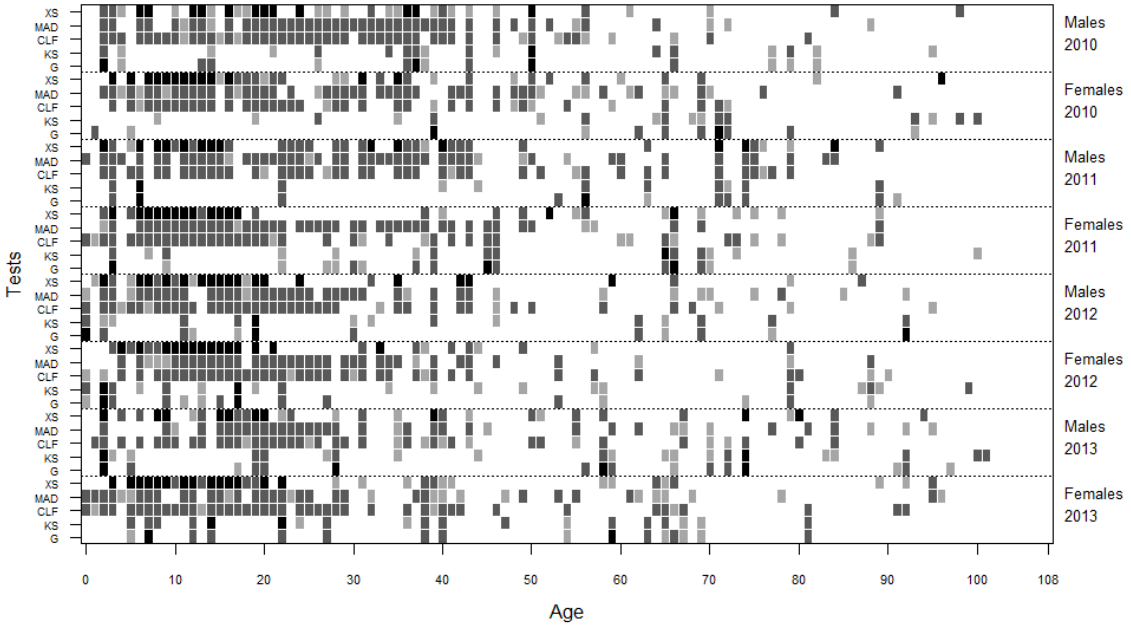

Figure 5-S. Uniform hypothesis tests by gender and age in Comunitat Valenciana population for immigrant events occurring in 2010-2013 Lexis cells (squares). Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.10 significant levels. MAD denotes the MAD test, CLF the Cressie–Loosmore–Ford test, XS the spatial Chi-squared goodness-of fit test, KS the Kolmogorov–Smirnov test and G the Geometric test. The three first tests are spatial tests and check CSR as null hypothesis after a representation of events in the Lexis space. KS and G tests are functional nonparametric tests and check whether the empirical distributions of exposed-to-risk times are compatible with the assumed probability distribution; which for squares is equivalent to assume that the density function of the random τ variable that measures the time lived with completed age x before immigrating in year t is $f(\tau) = 1$, for $0 \leq \tau < 1$.

EMIGRANT UNIFORM HYPOTHESIS TESTS

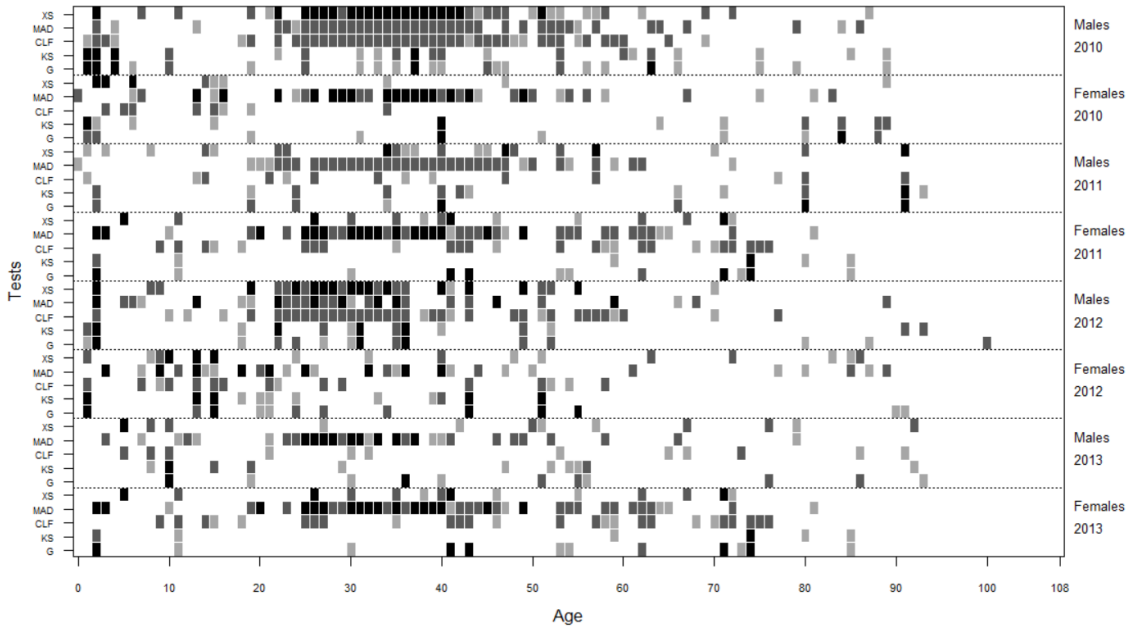

Figure 6-S. Uniform hypothesis tests by gender and age in Comunitat Valenciana population for emigrant events occurring in 2010-2013 Lexis upper triangles. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.10 significant levels. MAD denotes the MAD test, CLF the Cressie–Loosmore–Ford test, XS the spatial Chi-squared goodness-of fit test, KS the Kolmogorov–Smirnov test and G the Geometric test. The three first tests are spatial tests and check CSR as null hypothesis after a representation of events in the Lexis space. KS and G tests are functional nonparametric tests and check whether the empirical distributions of non-exposed-to-risk times are compatible with the assumed probability distribution.

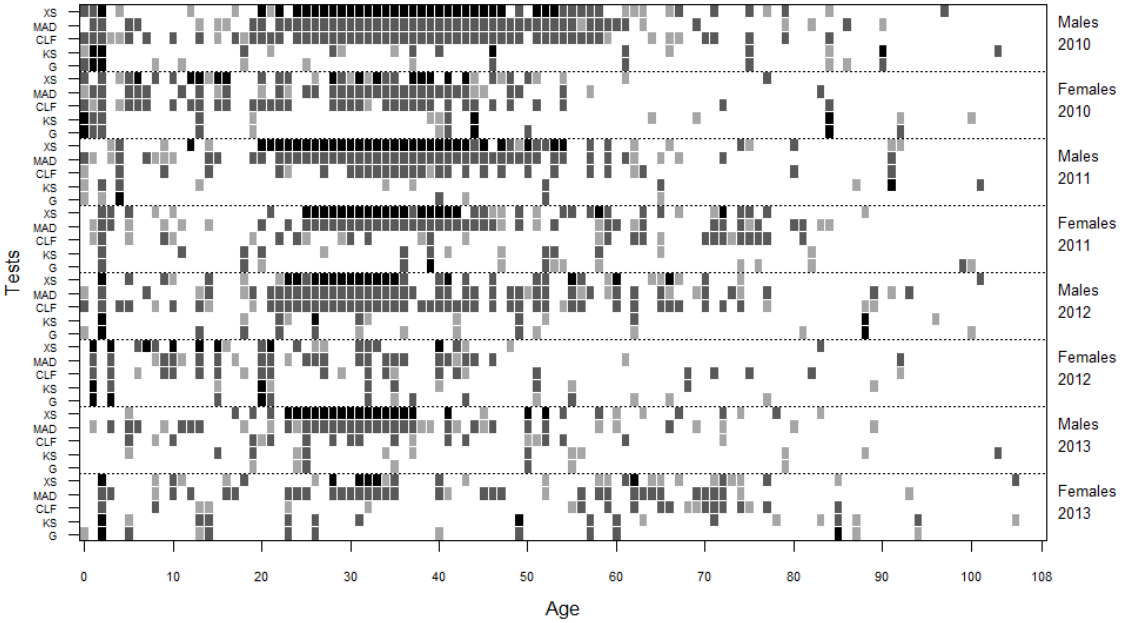

Figure 7-S. Uniform hypothesis tests by gender and age in Comunitat Valenciana population for emigrant events occurring in 2010-2013 Lexis cells (squares). Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.10 significant levels. MAD denotes the MAD test, CLF the Cressie–Loosmore–Ford test, XS the spatial Chi-squared goodness-of fit test, KS the Kolmogorov–Smirnov test and G the Geometric test. The three first tests are spatial tests and check CSR

as null hypothesis after a representation of events in the Lexis space. KS and G tests are functional nonparametric tests and check whether the empirical distributions of exposed-to-risk times are compatible with the assumed probability distribution; which for squares is equivalent to assume that the density function of the random τ variable that measures the time lived with completed age x before emigrating in year t is f(τ) = 1, for 0 ≤ τ < 1.

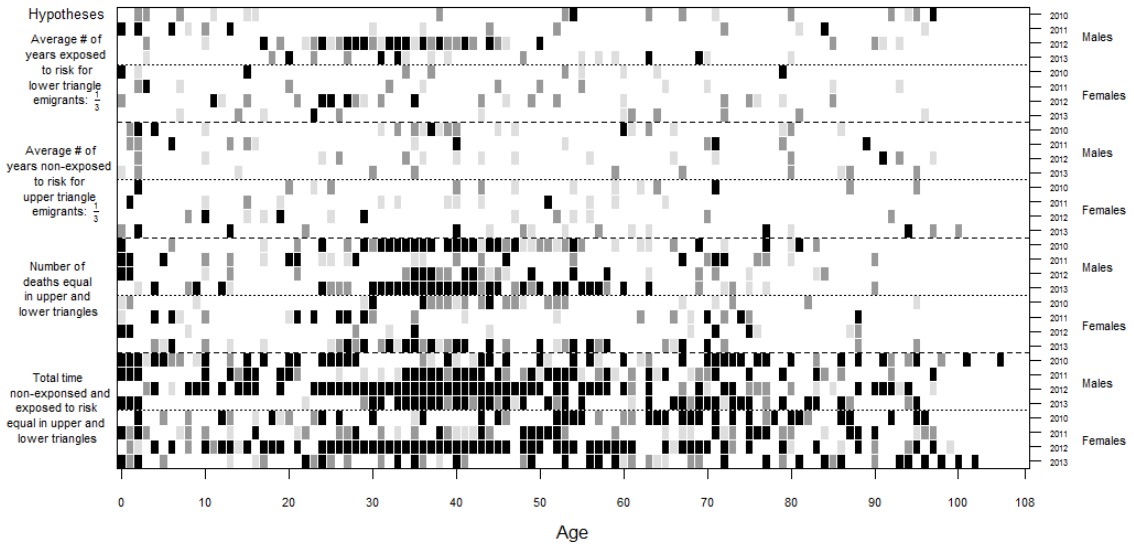

Figure 8-S. Parametric hypothesis tests, by gender and age for 2010–2013 Comunitat Valenciana population, corresponding to the concreteness in equations (1) and (2) of the hypotheses of uniform distribution of emigrants. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.10 significant levels. Two-side mean t-student tests are used to test the first two blocks of hypotheses, two-side binomial tests with $p = 0.5$ to assess the third block and a test based on the bootstrap empirical distribution of diferences to gauge the last block of hypotheses.

OTHER UNIFORM HYPOTHESIS TESTS

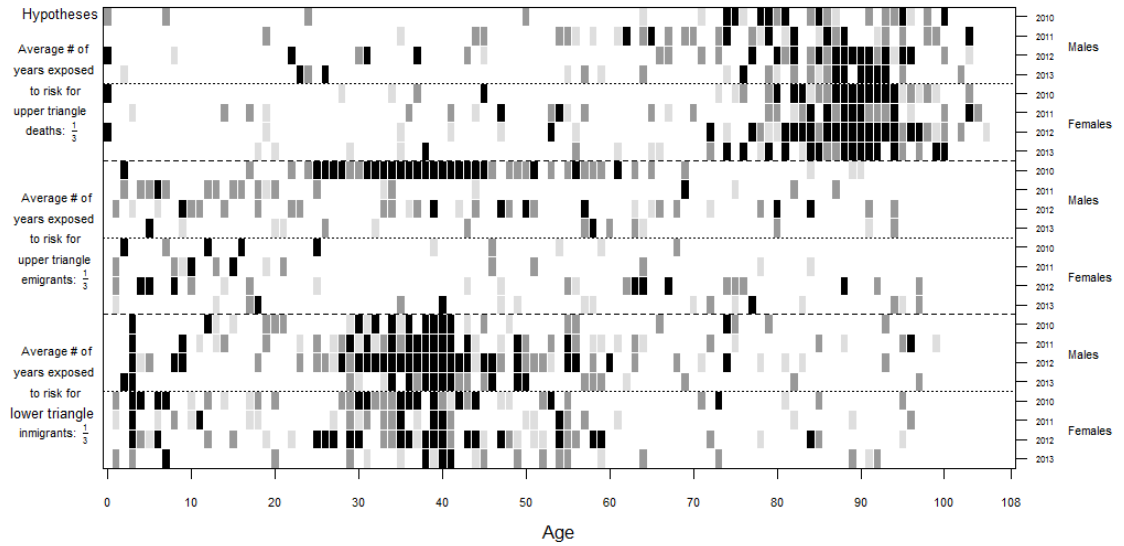

Figure 9-S. Rest of parametric hypothesis tests, by gender and age for 2010–2013 Comunitat Valenciana population, corresponding to the concreteness of the hypotheses of uniform distribution of deaths and migrants in terms of 'person-years' exposed-at-risk. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.10 significant levels. Two-side mean t-student tests are used to test the hypotheses.

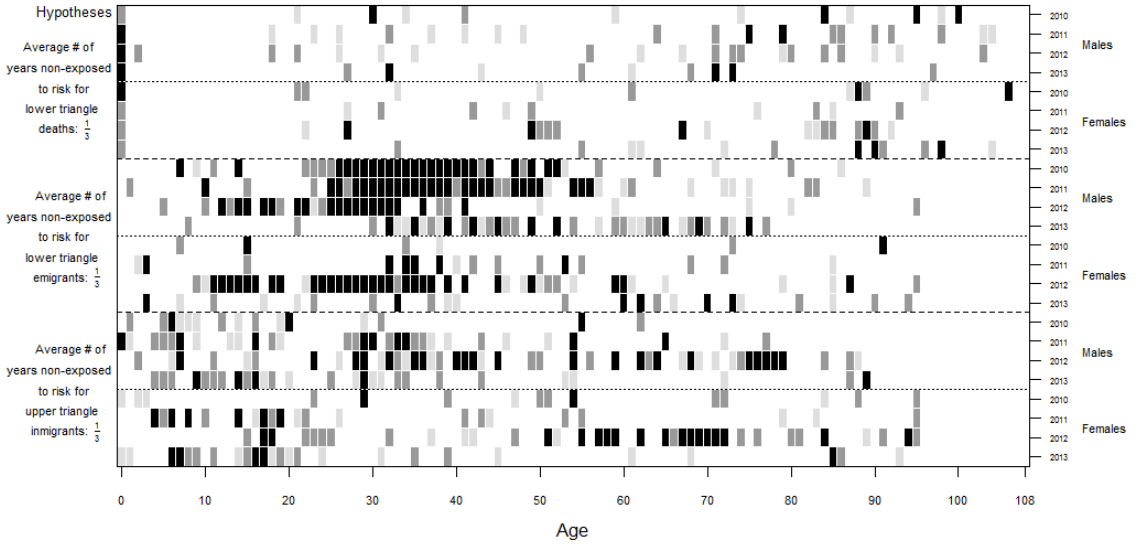

Figure 10-S. Rest of parametric hypothesis tests, by gender and age for 2010–2013 Comunitat Valenciana population, corresponding to the concreteness of the hypotheses of uniform distribution of deaths and migrants in terms of 'person-years' non-exposed-at-risk. Black, dark grey and light grey colours indicate rejection of null hypotheses at, respectively, 0.01, 0.05 and 0.10 significant levels. Two-side mean t-student tests are used to test the hypotheses.

TEMPORAL DISTRIBUTION OF BIRTHS

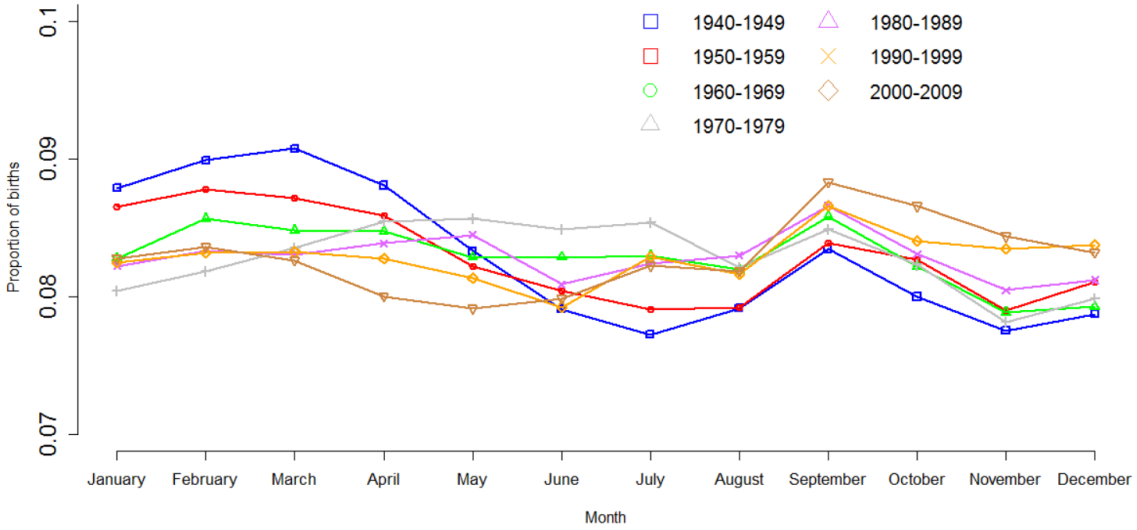

Figure 11-S. Monthly distribution of births in Comunitat Valenciana (Spain) in some decades. The monthly proportions have been standardized to account for the different number of days that each month has.

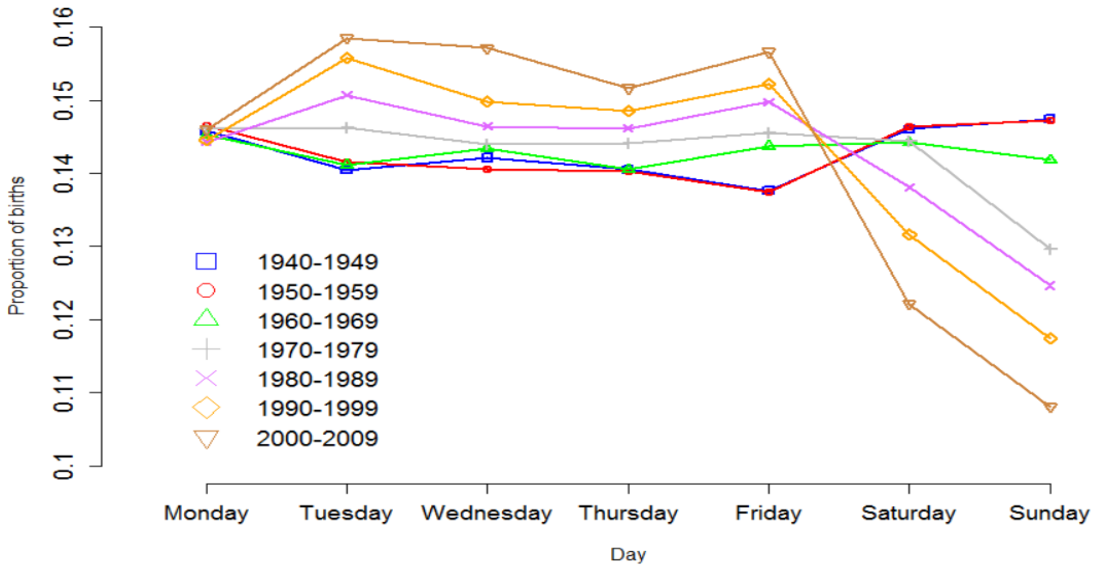

Figure 12-S. Weekly distribution of births in Comunitat Valenciana (Spain) in some decades.

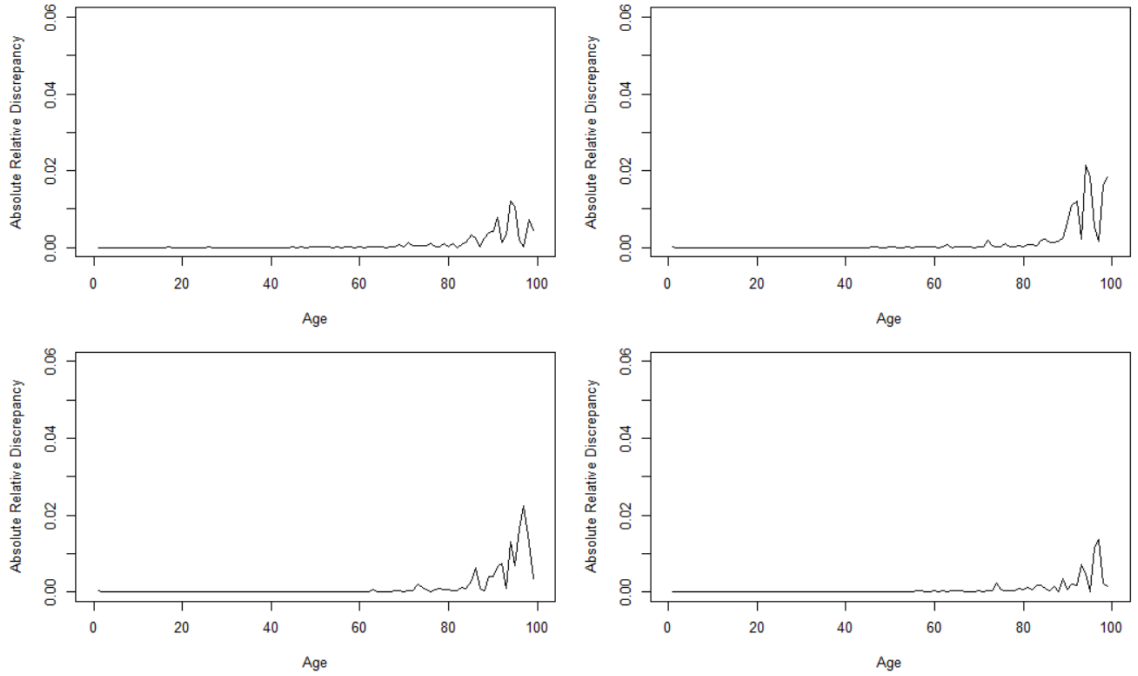

Figure 13-S. Absolute relative discrepancies between the estimated rates of death, m_x, of either CP_UD_UB (closed population and uniform distribution of deaths and births) or OP_UD_UM_UB (open population and uniform distribution of deaths, migrants and births) scenarios and of CP_NUD_UB (closed population with no hypothesis about distribution of deaths and uniform distribution of births) scenario for men, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

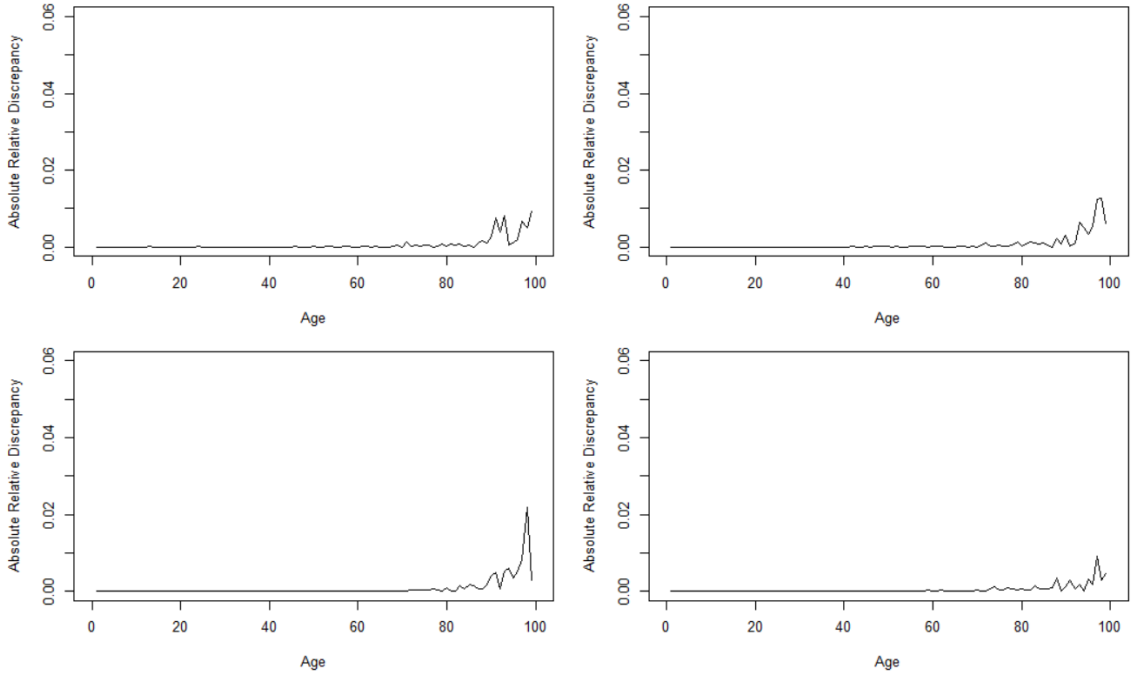

Figure 14-S. Absolute relative discrepancies between the estimated rates of death, m_x, of either CP_UD_UB (closed population and uniform distribution of deaths and births) or OP_UD_UM_UB (open population and uniform distribution of deaths, migrants and births) scenarios and of CP_NUD_UB (closed population with no hypothesis about distribution of deaths and uniform distribution of births) scenario

for women, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

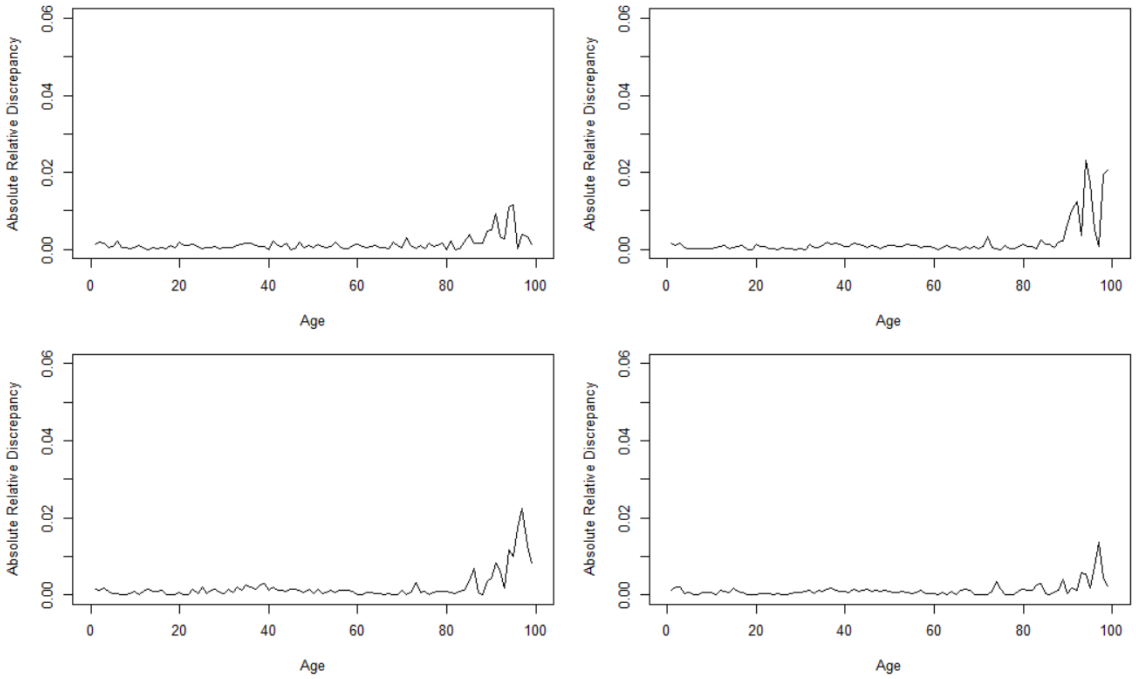

Figure 15-S. Absolute relative discrepancies between the estimated rates of death, m_x, of either CP_UD_UB (closed population and uniform distribution of deaths and births) or OP_UD_UM_UB (open population and uniform distribution of deaths, migrants and births) scenarios and of OP_NUD_NUM_UB (open population with no hypotheses about distribution of deaths and migrants and uniform distribution of births) scenario for men, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

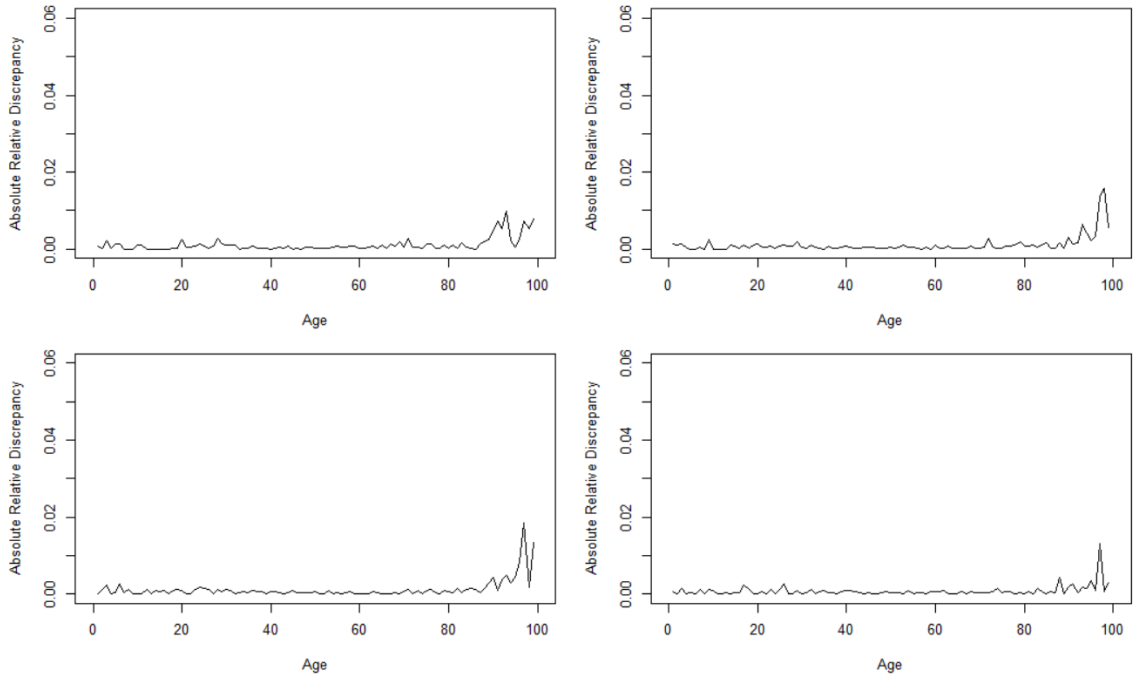

Figure 16-S. Absolute relative discrepancies between the estimated rates of death, m_x, of either CP_UD_UB (closed population and uniform distribution of deaths and births) or OP_UD_UM_UB (open population and uniform distribution of deaths, migrants and births) scenarios and of OP_NUD_NUM_UB (open population with no hypotheses about distribution of deaths and migrants and uniform distribution

of births) scenario for women, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

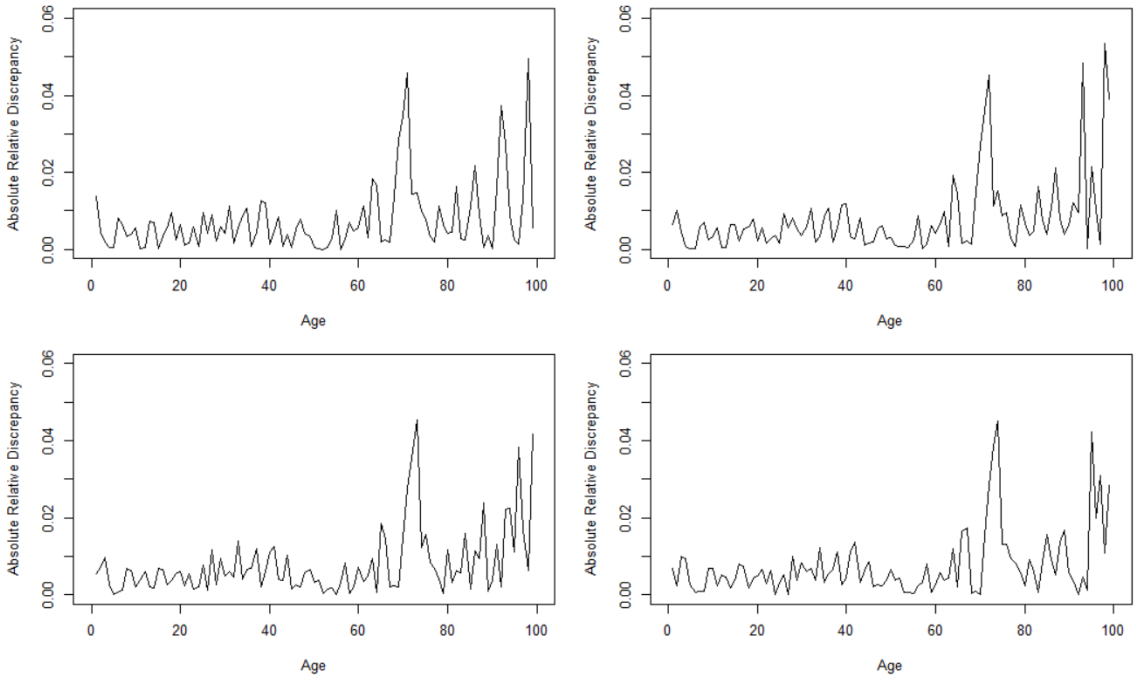

Figure 17-S. Absolute relative discrepancies between the estimated rates of death, m_x, of either CP_UD_UB (closed population and uniform distribution of deaths and births) or OP_UD_UM_UB (open population and uniform distribution of deaths, migrants and births) scenarios and of OP_NUD_NUM_NUB (open population with no hypotheses about distribution of deaths, migrants and births) scenario for men, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

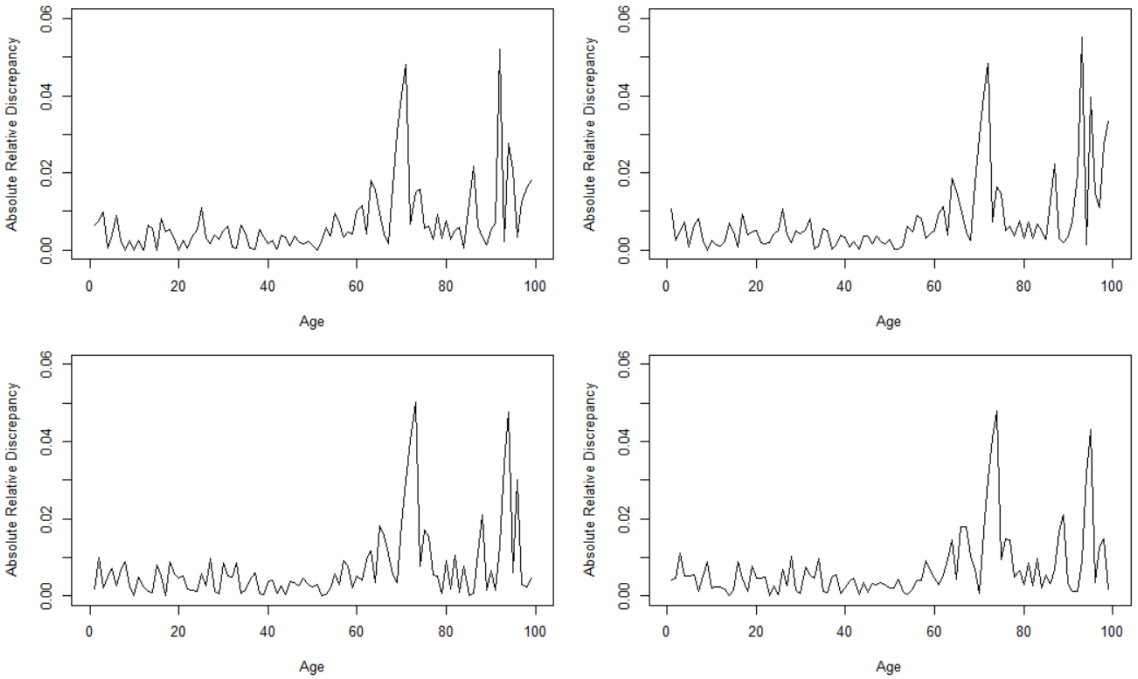

Figure 18-S. Absolute relative discrepancies between the estimated rates of death, m_x, of either CP_UD_UB (closed population and uniform distribution of deaths and births) or OP_UD_UM_UB (open population and uniform distribution of deaths, migrants and births) scenarios and of OP_NUD_NUM_NUB (open population with no hypotheses about distribution of deaths, migrants and births) scenario for

women, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

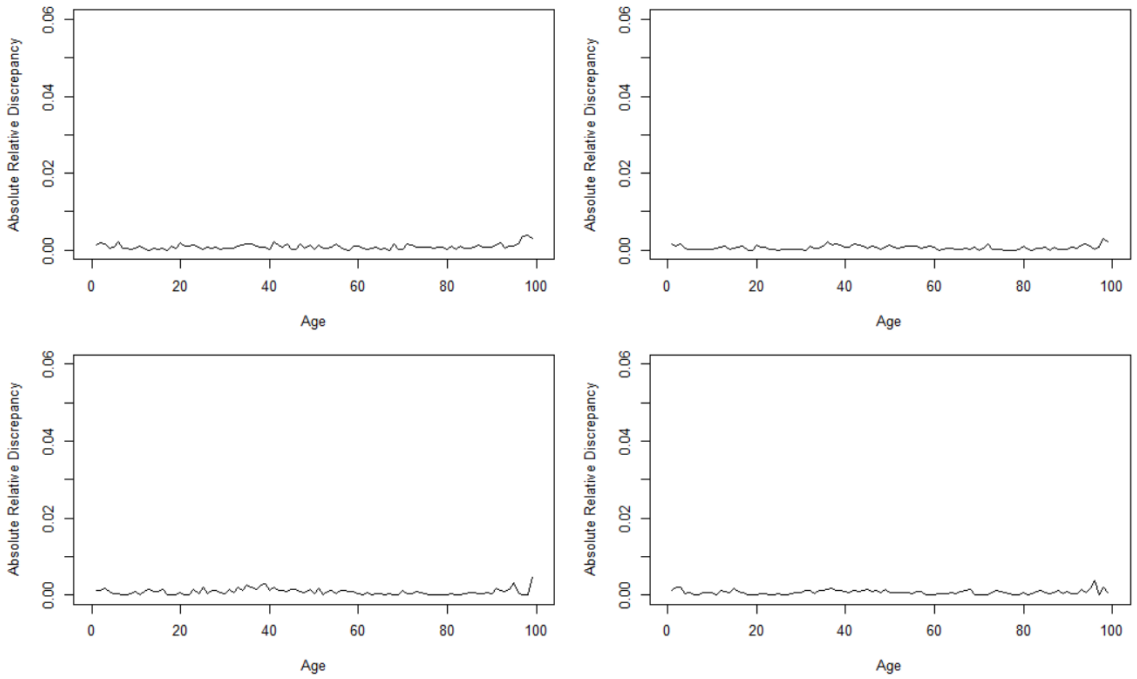

Figure 19-S. Absolute relative discrepancies between the estimated rates of death, m_x, of CP_NUD_UB (closed population with no hypothesis about distribution of deaths and uniform distribution of births) scenario and of OP_NUD_NUM_UB (open population with no hypotheses about distribution of deaths and migrants and uniform distribution of births) scenario for men, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

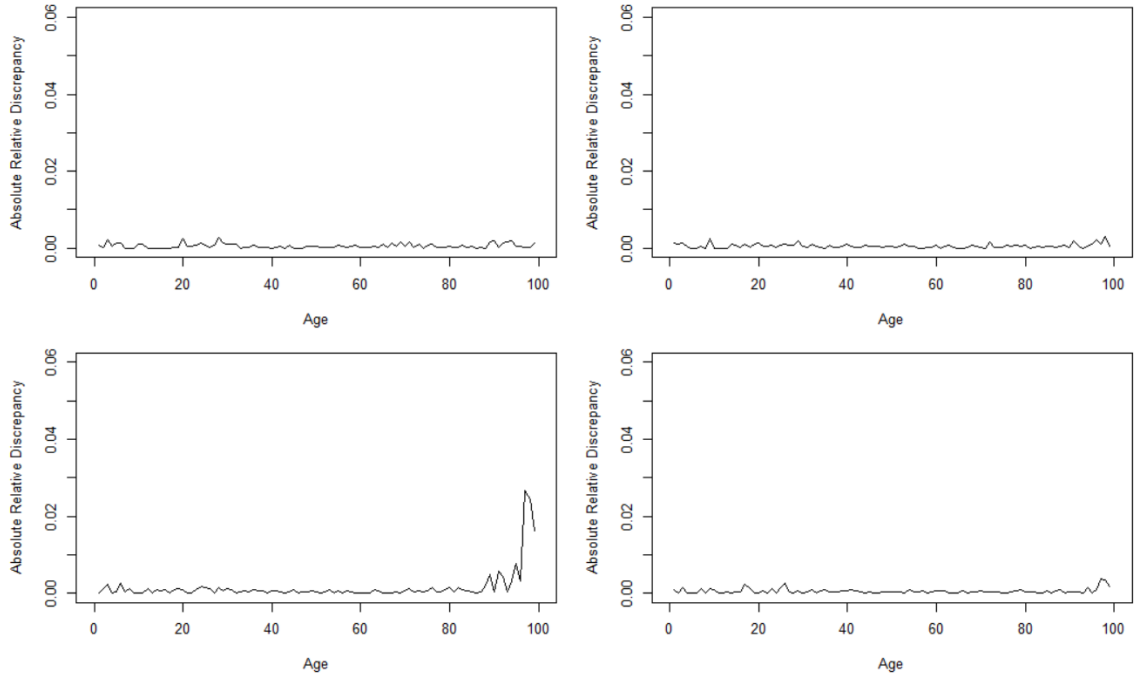

Figure 20-S. Absolute relative discrepancies between the estimated rates of death, m_x, of CP_NUD_UB (Closed population with no hypothesis about distribution of deaths and uniform distribution of births) scenario and of OP_NUD_NUM_UB (open population with no hypotheses about distribution of deaths and migrants and uniform distribution of births) scenario for women, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

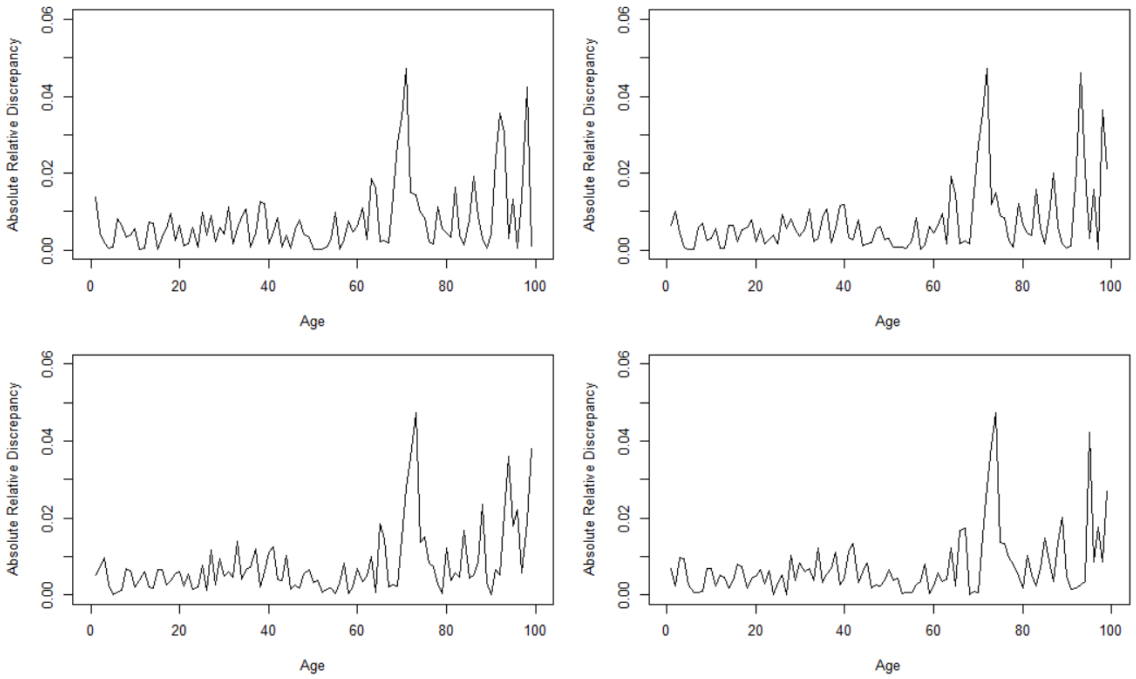

Figure 21-S. Absolute relative discrepancies between the estimated rates of death, m_x, of CP_NUD_UB (Closed population with no hypothesis about distribution of deaths and uniform distribution of births) scenario and of OP_NUD_NUM_NUB (open population with no hypotheses about distribution of deaths, migrants and births) scenario for men, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

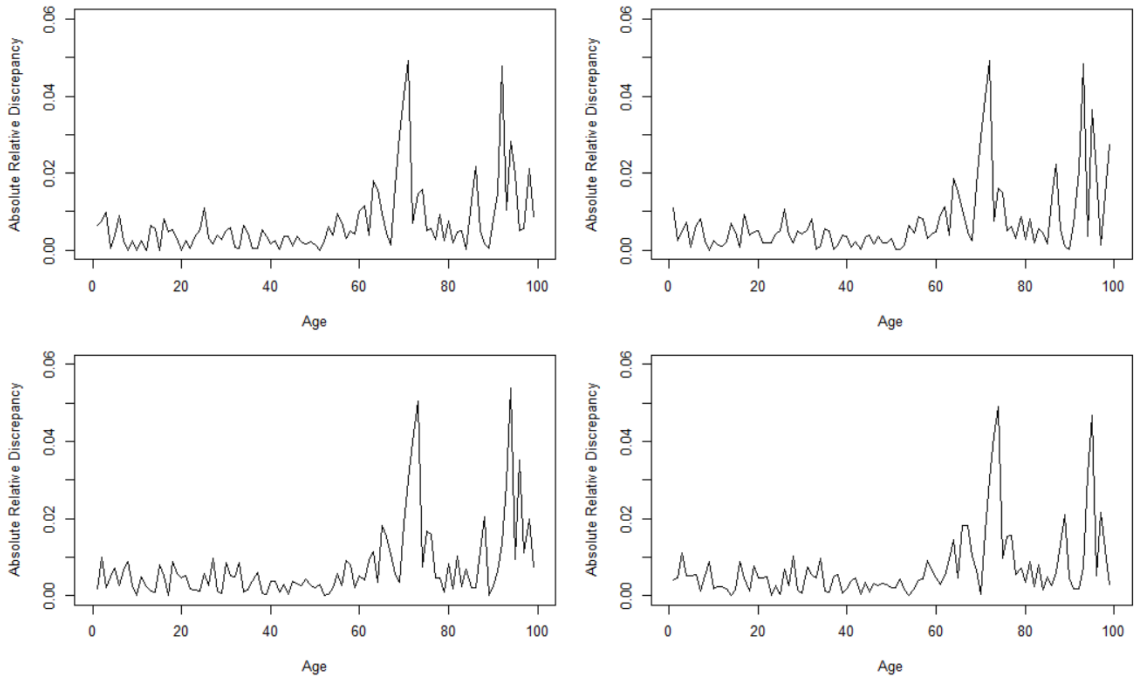

Figure 22-S. Absolute relative discrepancies between the estimated rates of death, m_x, of CP_NUD_UB (Closed population with no hypothesis about distribution of deaths and uniform distribution of births) scenario and of OP_NUD_NUM_NUB (open population with no hypotheses about distribution of deaths, migrants and births) scenario for women, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

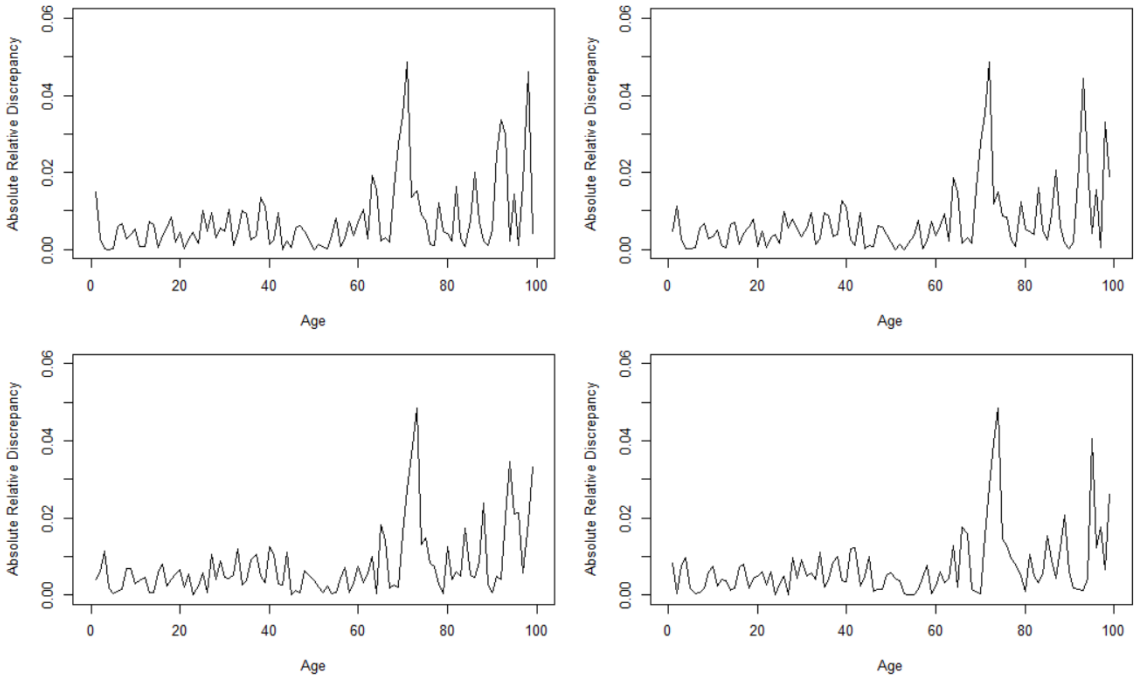

Figure 23-S. Absolute relative discrepancies between the estimated rates of death, m_x, of OP_NUD_NUM_UB (open population with no hypotheses about distribution of deaths and migrants and uniform distribution of births) scenario and of OP_NUD_NUM_NUB (open population with no hypotheses about distribution of deaths, migrants and births) scenario for men, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

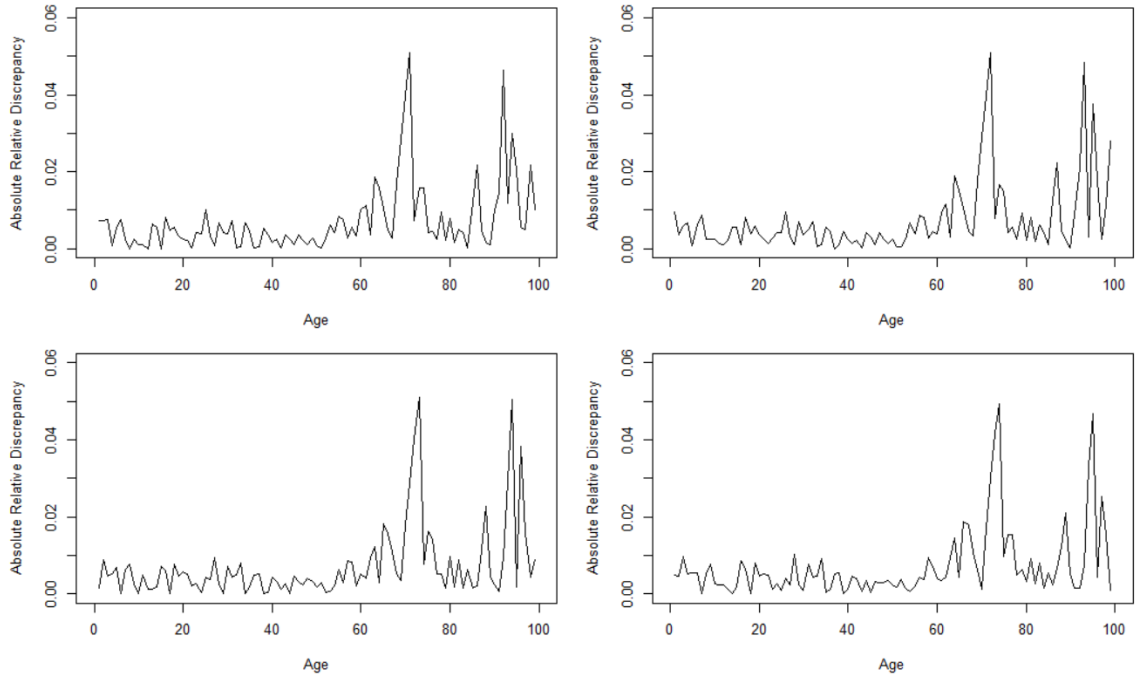

Figure 24-S. Absolute relative discrepancies between the estimated rates of death, m_x, of OP_NUD_NUM_UB (open population with no hypotheses about distribution of deaths and migrants and uniform distribution of births) scenario and of OP_NUD_NUM_NUB (open population with no hypotheses about distribution of deaths, migrants and births) scenario for women, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

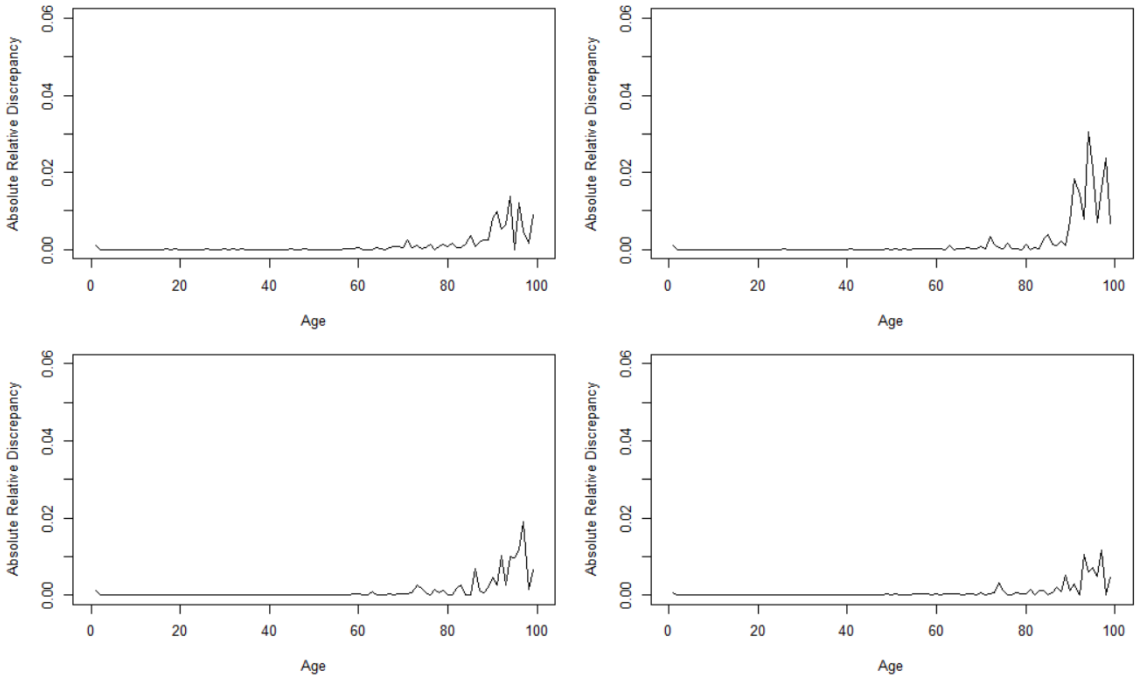

Figure 25-S. Absolute relative discrepancies between the probabilities of death, q_x, of either CP_UD_UB (closed population and uniform distribution of deaths and births) or OP_UD_UM_UB (open population and uniform distribution of deaths, migrants and births) scenarios and of CP_NUD_UB (closed population with no hypothesis about distribution of deaths and uniform distribution of births) scenario for men, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

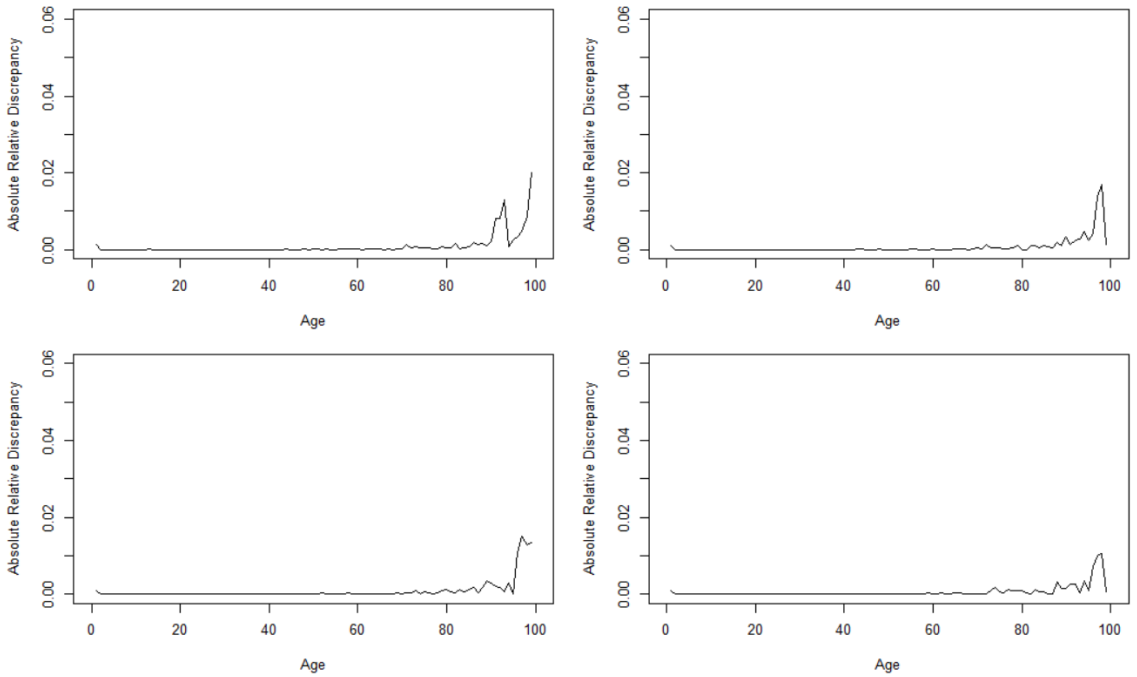

Figure 26-S. Absolute relative discrepancies between the probabilities of death, q_x, of either CP_UD_UB (closed population and uniform distribution of deaths and births) or OP_UD_UM_UB (open population and uniform distribution of deaths, migrants and births) scenarios and of CP_NUD_UB (closed population with no hypothesis about distribution of deaths and uniform distribution of births) scenario for women, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

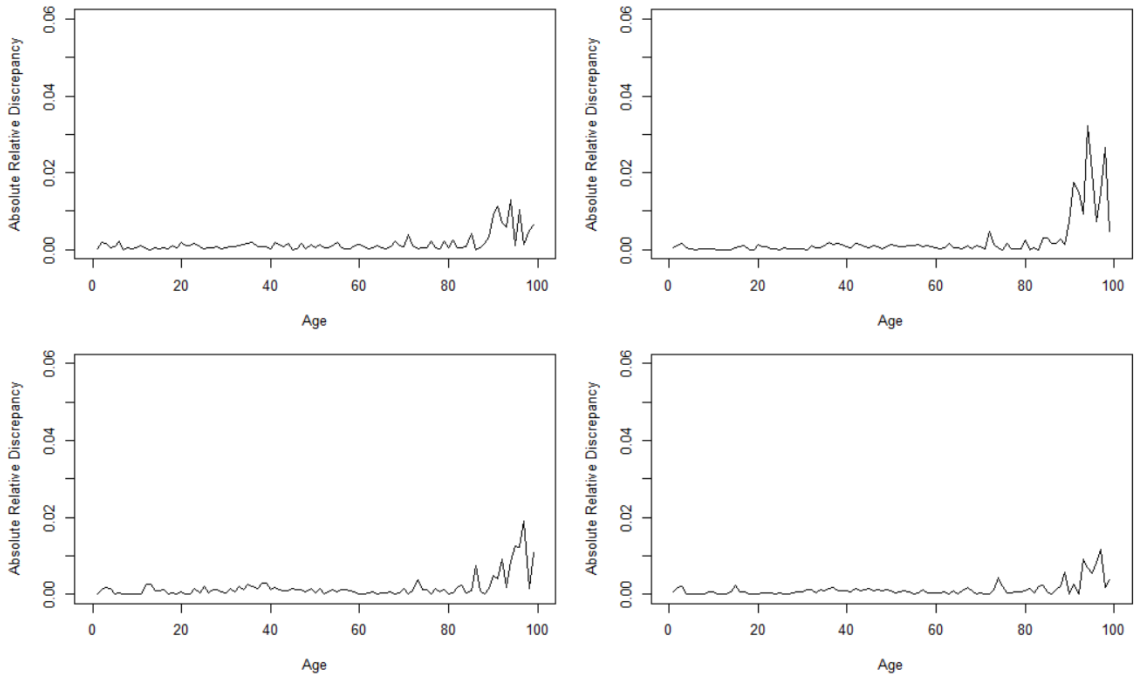

Figure 27-S. Absolute relative discrepancies between the probabilities of death, q_x, of either CP_UD_UB (closed population and uniform distribution of deaths and births) or OP_UD_UM_UB (open population and uniform distribution of deaths, migrants and births) scenarios and of OP_NUD_NUM_UB (open population with no hypotheses about distribution of deaths and migrants and uniform distribution of births) scenario for men, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

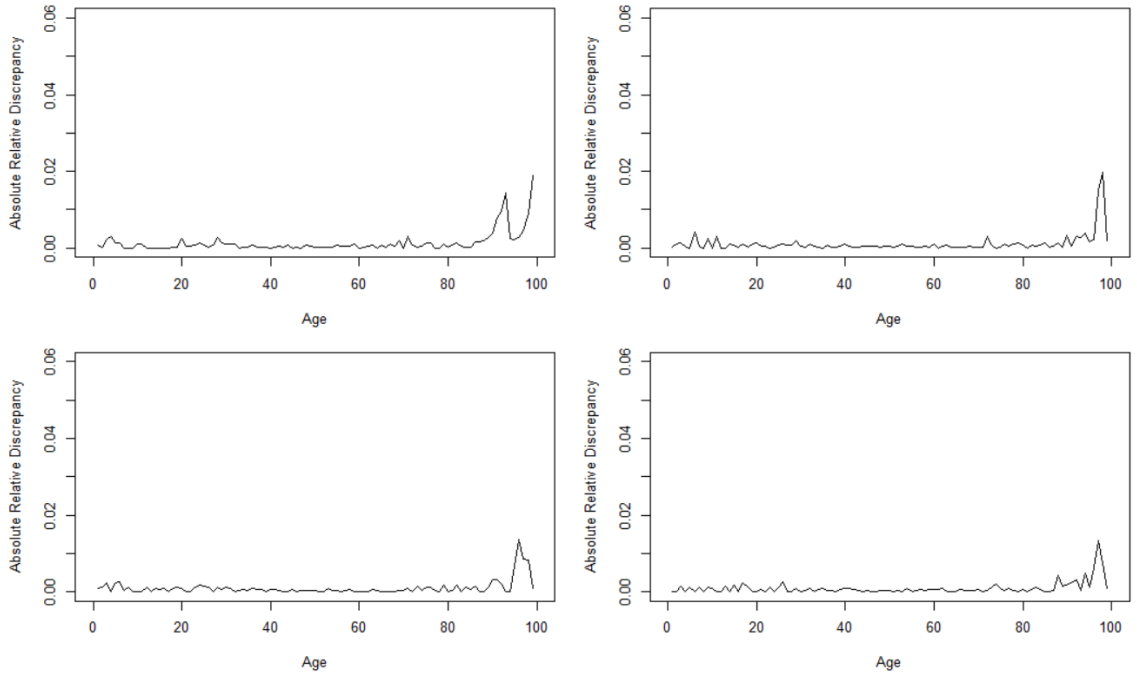

Figure 28-S. Absolute relative discrepancies between the probabilities of death, q_x, of either CP_UD_UB (closed population and uniform distribution of deaths and births) or OP_UD_UM_UB (open population and uniform distribution of deaths, migrants and births) scenarios and of OP_NUD_NUM_UB (open population with no hypotheses about distribution of deaths and migrants and uniform distribution of births) scenario for women, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

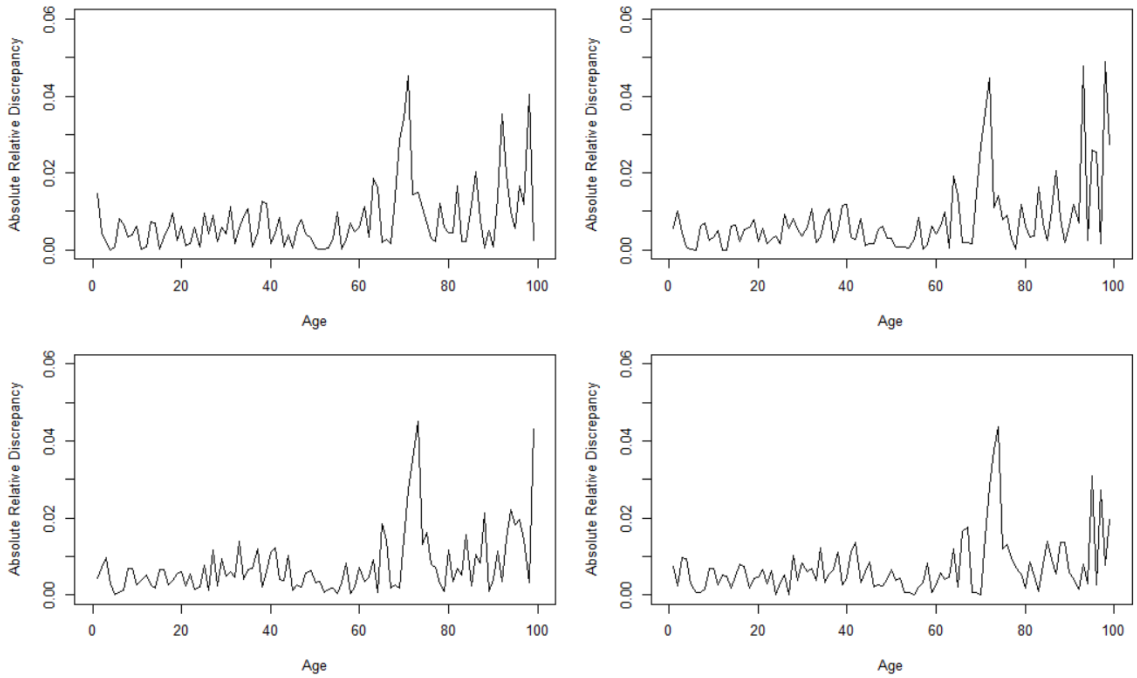

Figure 29-S. Absolute relative discrepancies between the probabilities of death, q_x, of either CP_UD_UB (closed population and uniform distribution of deaths and births) or OP_UD_UM_UB (open population and uniform distribution of deaths, migrants and births) scenarios and of OP_NUD_NUM_NUB (open population with no hypotheses about distribution of deaths, migrants and births) scenario for men, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

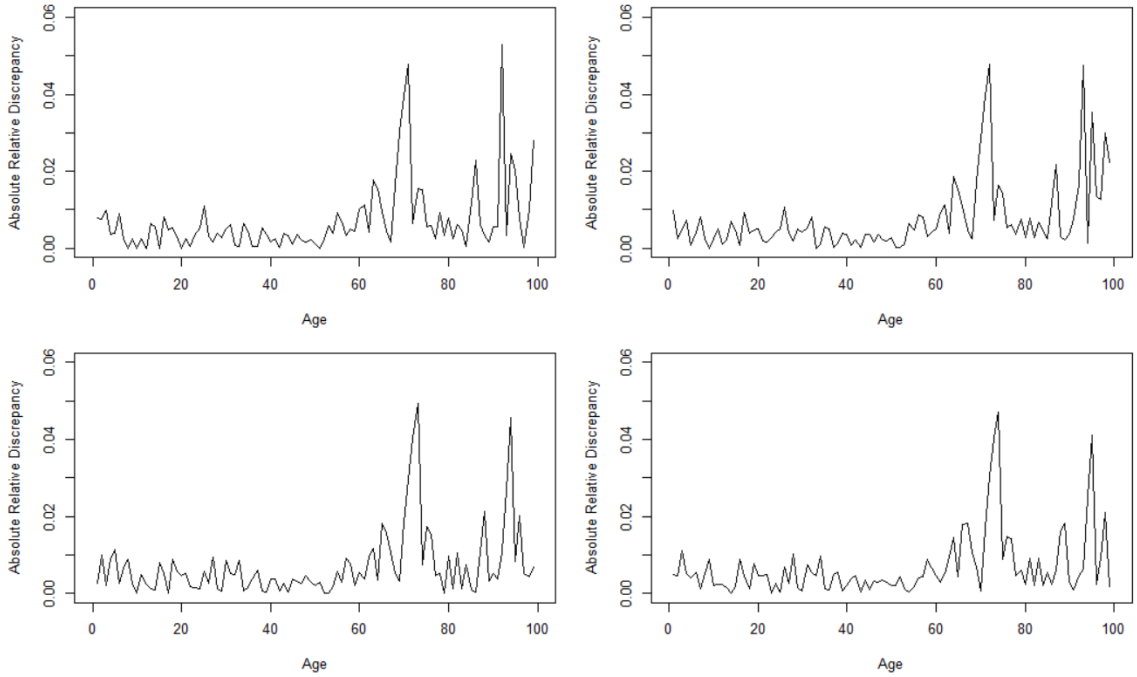

Figure 30-S. Absolute relative discrepancies between the probabilities of death, q_x, of either CP_UD_UB (closed population and uniform distribution of deaths and births) or OP_UD_UM_UB (open population and uniform distribution of deaths, migrants and births) scenarios and of OP_NUD_NUM_NUB (open population with no hypotheses about distribution of deaths, migrants and births) scenario for women, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

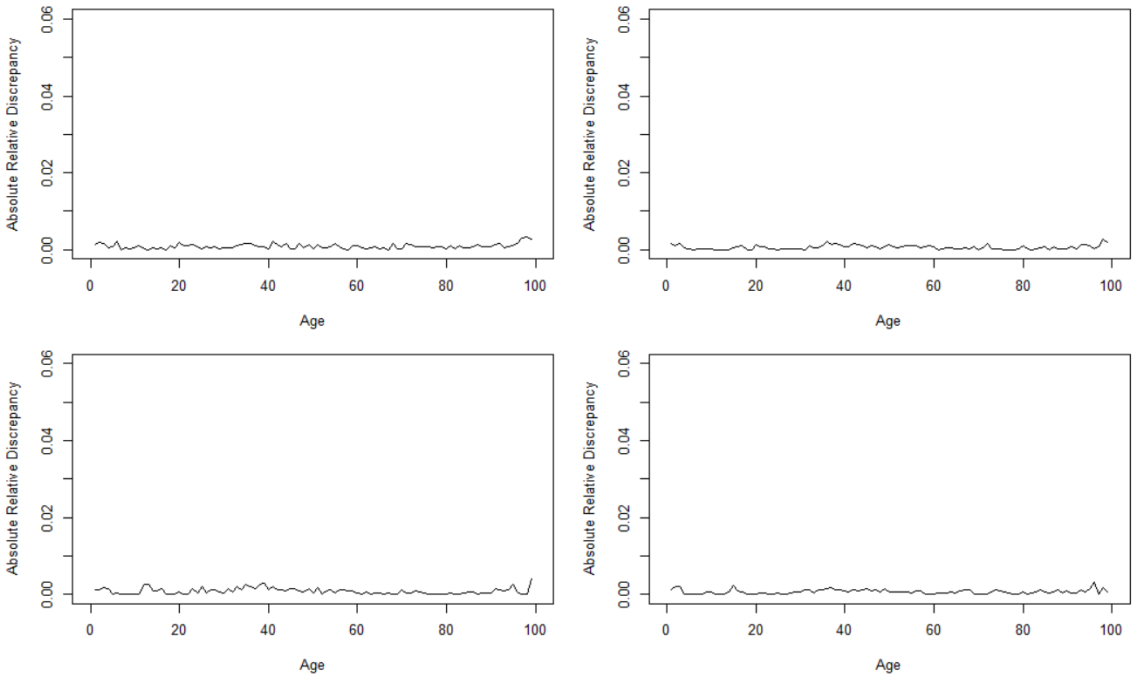

Figure 31-S. Absolute relative discrepancies between the probabilities of death, q_x, of CP_NUD_UB (closed population with no hypothesis about distribution of deaths and uniform distribution of births) scenario and of OP_NUD_NUM_UB (open population with no hypotheses about distribution of deaths and migrants and uniform distribution of births) scenario for men, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

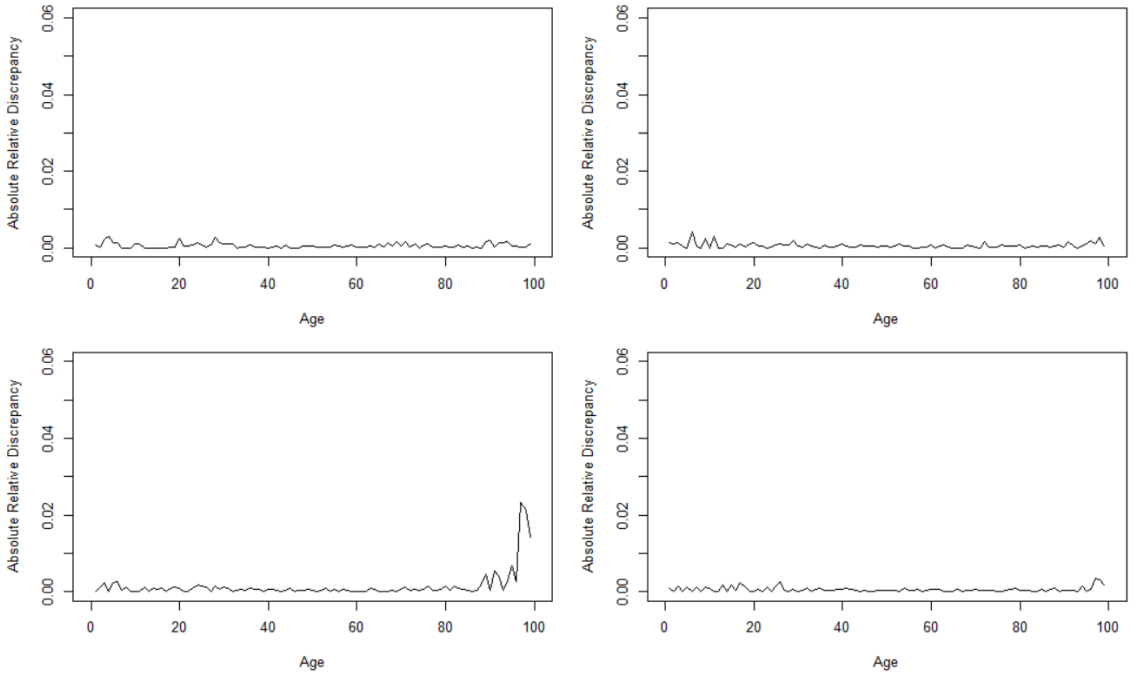

Figure 32-S. Absolute relative discrepancies between the probabilities of death, q_x, of CP_NUD_UB (closed population with no hypothesis about distribution of deaths and uniform distribution of births) scenario and of OP_NUD_NUM_NUB (open population with no hypotheses about distribution of deaths, migrants and births) scenario for women, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

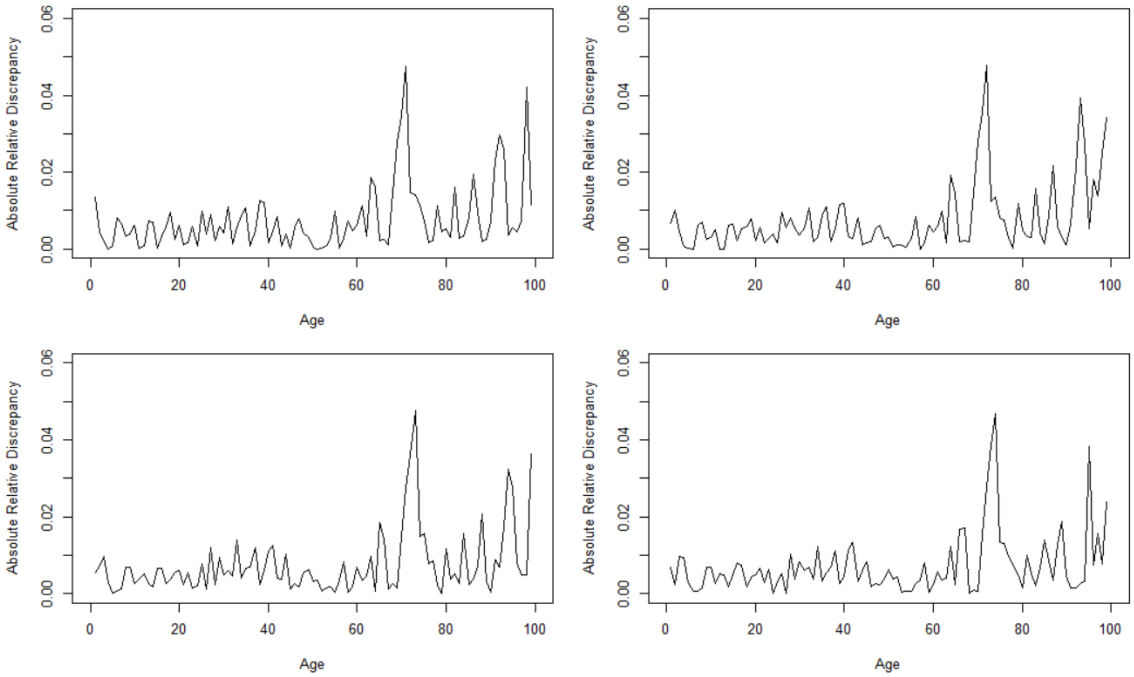

Figure 33-S. Absolute relative discrepancies between the probabilities of death, q_x, of CP_NUD_UB (Closed population with no hypothesis about distribution of deaths and uniform distribution of births) scenario and of OP_NUD_NUM_NUB (open population with no hypotheses about distribution of deaths, migrants and births) scenario for men, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

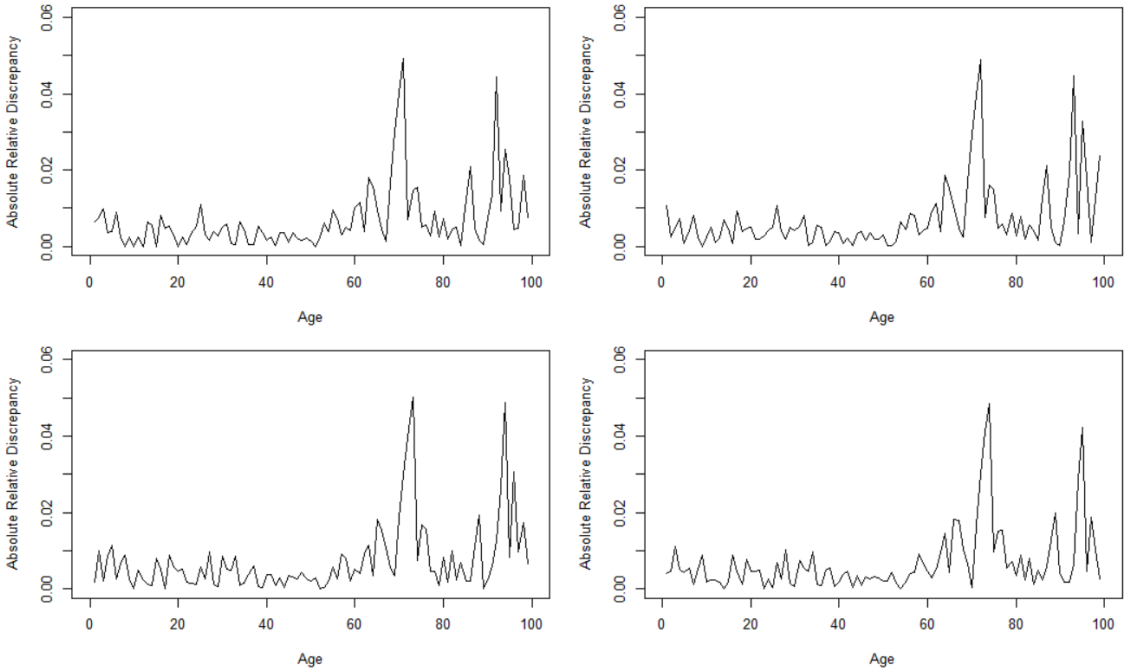

Figure 34-S. Absolute relative discrepancies between the probabilities of death, q_x, of CP_NUD_UB (Closed population with no hypothesis about distribution of deaths and uniform distribution of births) scenario and of OP_NUD_NUM_NUB (open population with no hypotheses about distribution of deaths, migrants and births) scenario for women, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

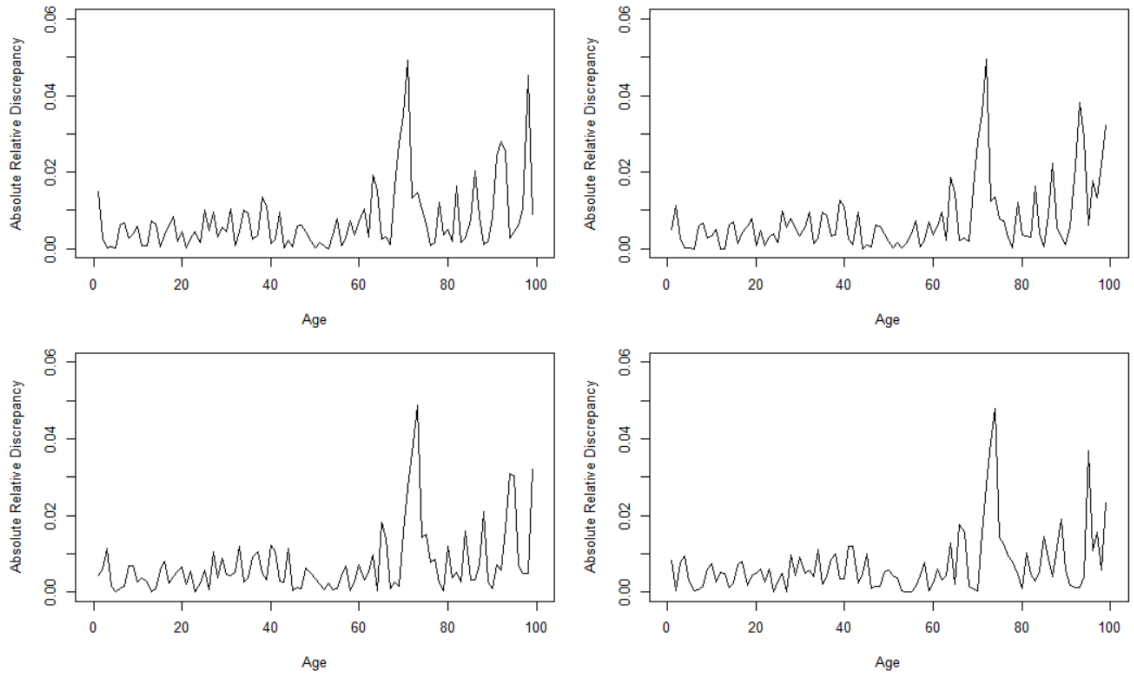

Figure 35-S. Absolute relative discrepancies between the probabilities of death, q_x, of OP_NUD_NUM_UB (open population with no hypotheses about distribution of deaths and migrants and uniform distribution of births) scenario and of OP_NUD_NUM_NUB (open population with no hypotheses about distribution of deaths, migrants and births) scenario for men, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

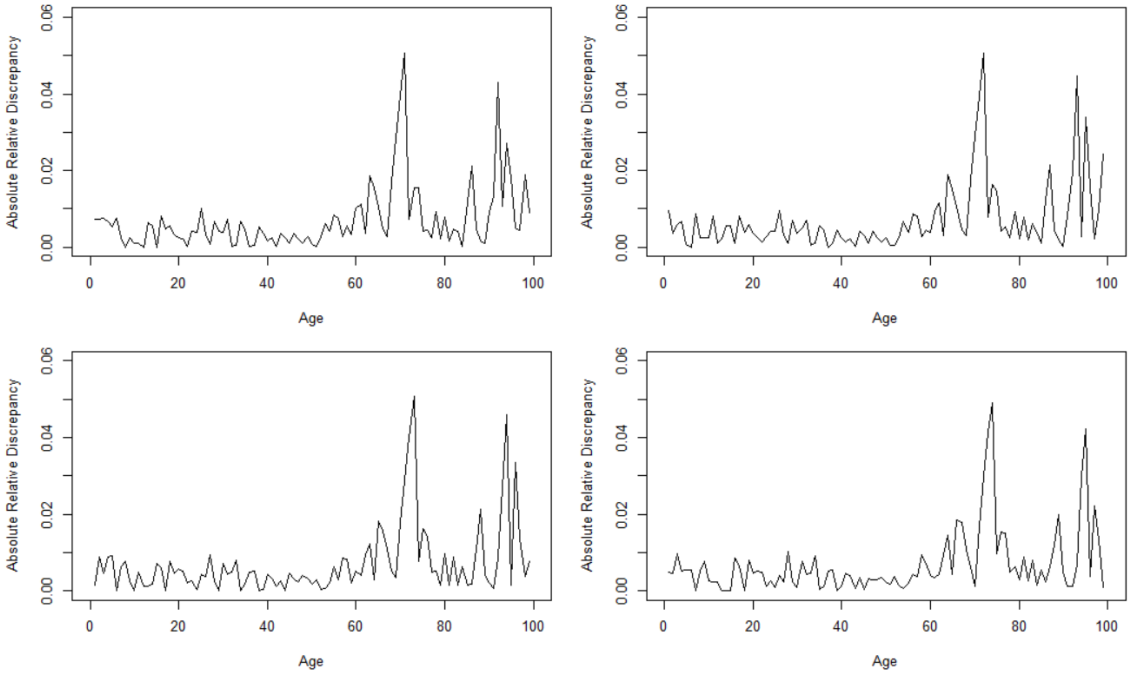

Figure 36-S. Absolute relative discrepancies between the probabilities of death, q_x, of OP_NUD_NUM_UB (open population with no hypotheses about distribution of deaths and migrants and uniform distribution of births) scenario and of OP_NUD_NUM_NUB (open population with no hypotheses about distribution of deaths, migrants and births) scenario for women, period 2010-2013. Left upper panel 2010, right upper panel 2011, left lower panel 2012 and right lower panel 2013.

ADDITIONAL FIGURES

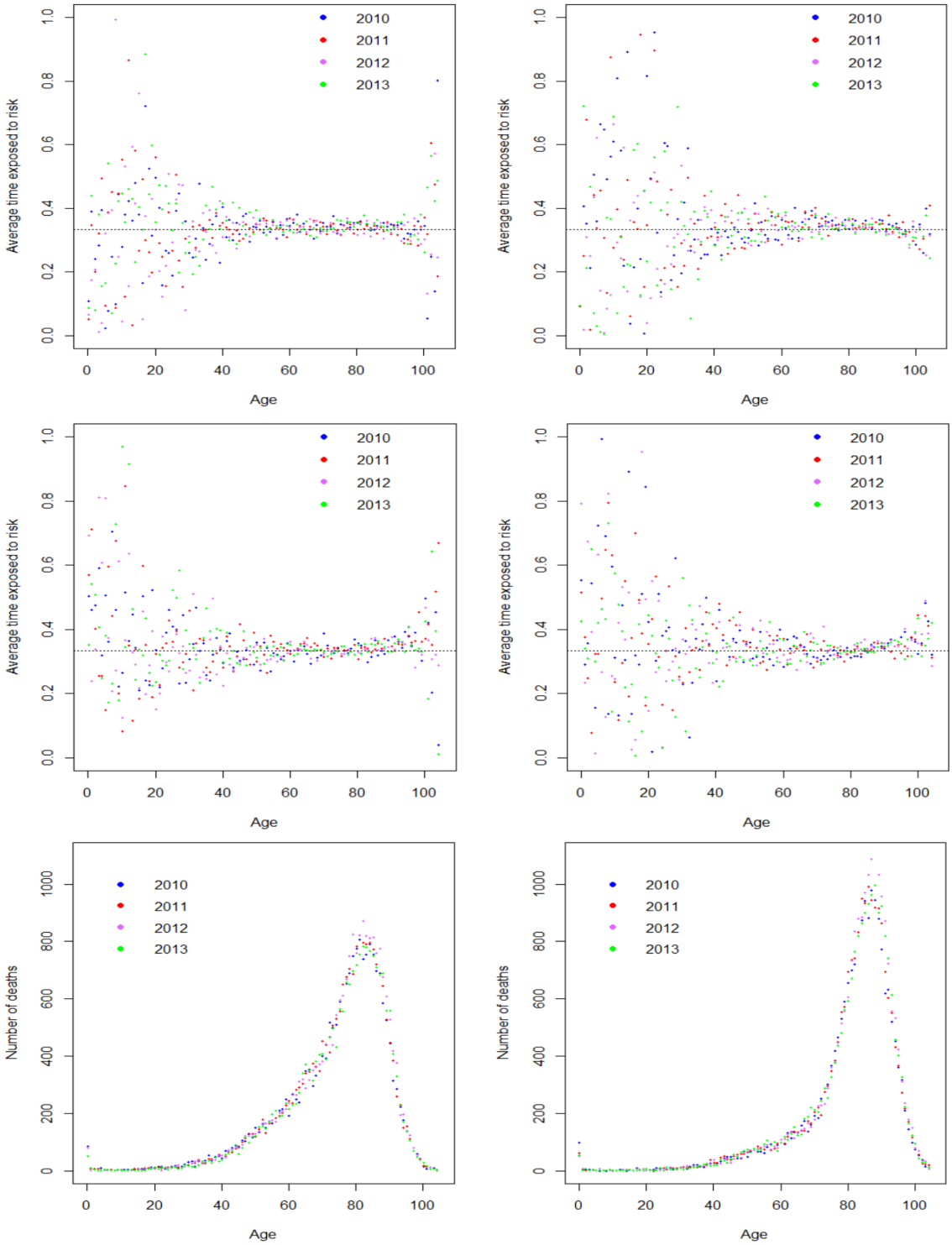

Figure 37-S. Average number of years exposed to risk of dying for deaths registered in Lexis lower triangles (upper panels) and lower triangles (middle panels) by age for males (left-panels) and females (right-panels). Lower panels: Number of deaths registered in Lexis squares by age for males (left-panel) and females (right-panel).

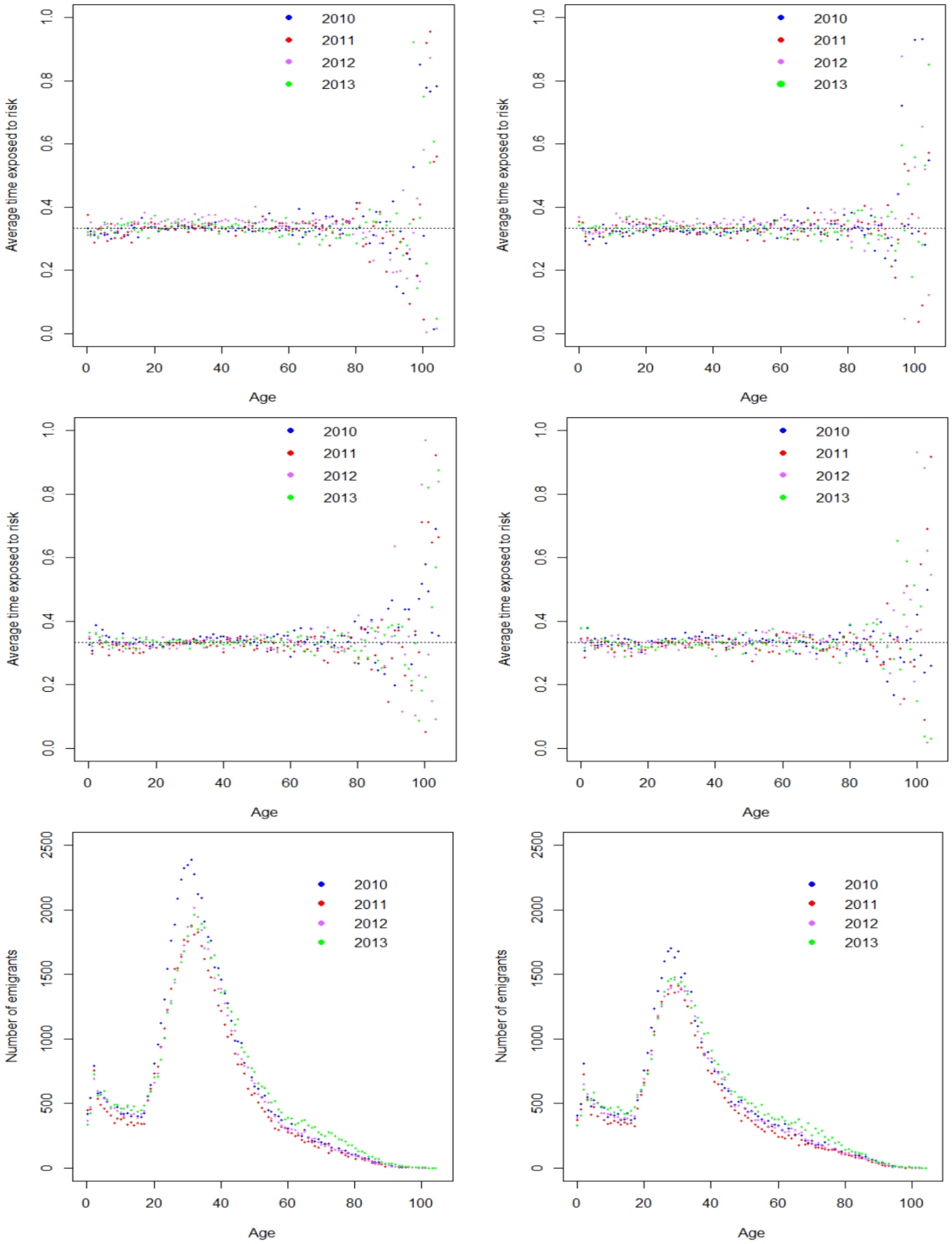

Figure 38-S. Average number of years exposed to risk of dying for emigrants located in Lexis lower triangles (upper panels) and lower triangles (middle panels) by age for males (left-panels) and females (right-panels). Lower panels: Number of emigrants registered in Lexis square by age for males (left-panel) and females (right-panel).

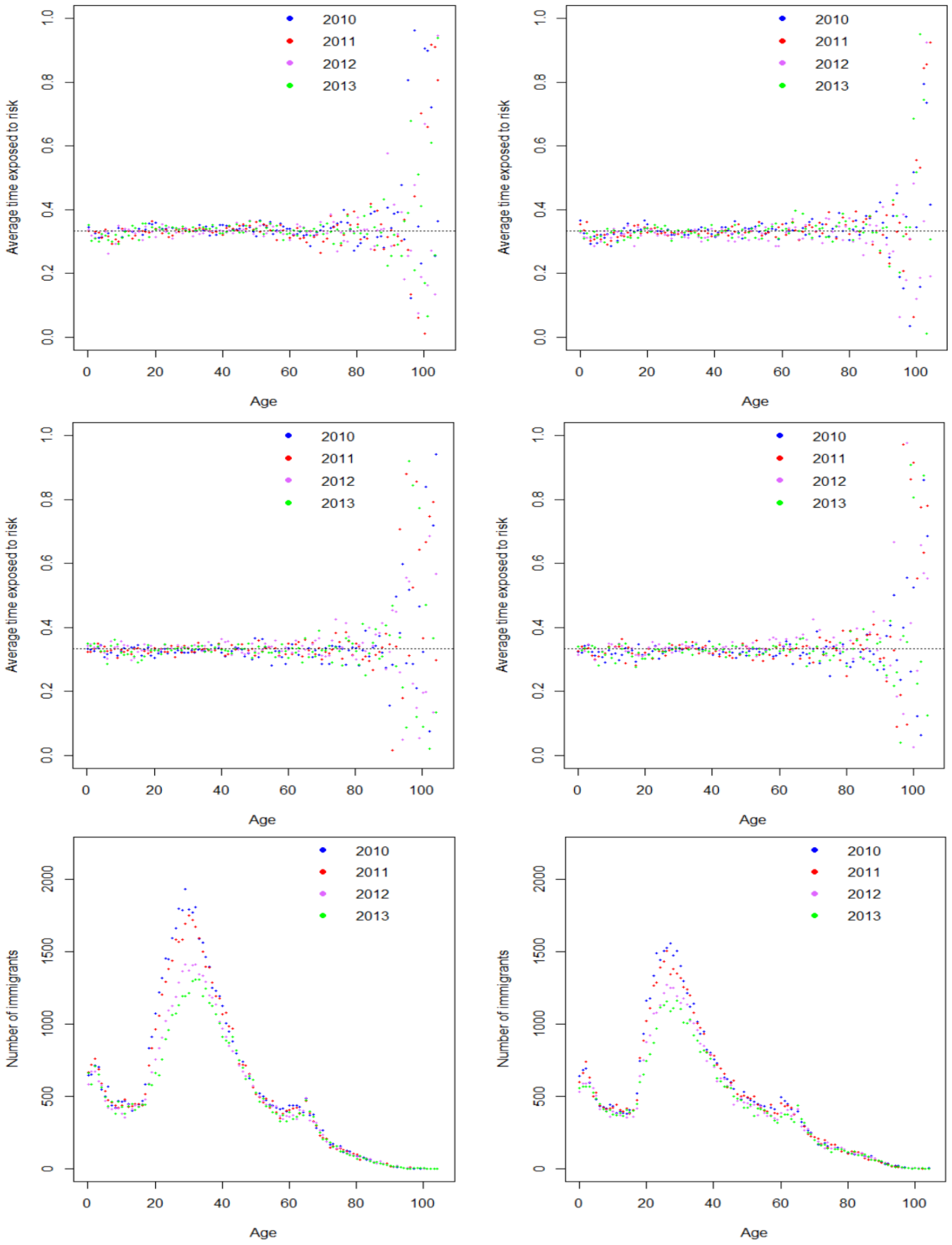

Figure 39-S. Average number of years exposed to risk of dying for immigrants located in Lexis lower triangles (upper panels) and lower triangles (middle panels) by age for males (left-panels) and females (right-panels). Lower panels: Number of immigrants registered in Lexis square by age for males (left-panel) and females (right-panel).

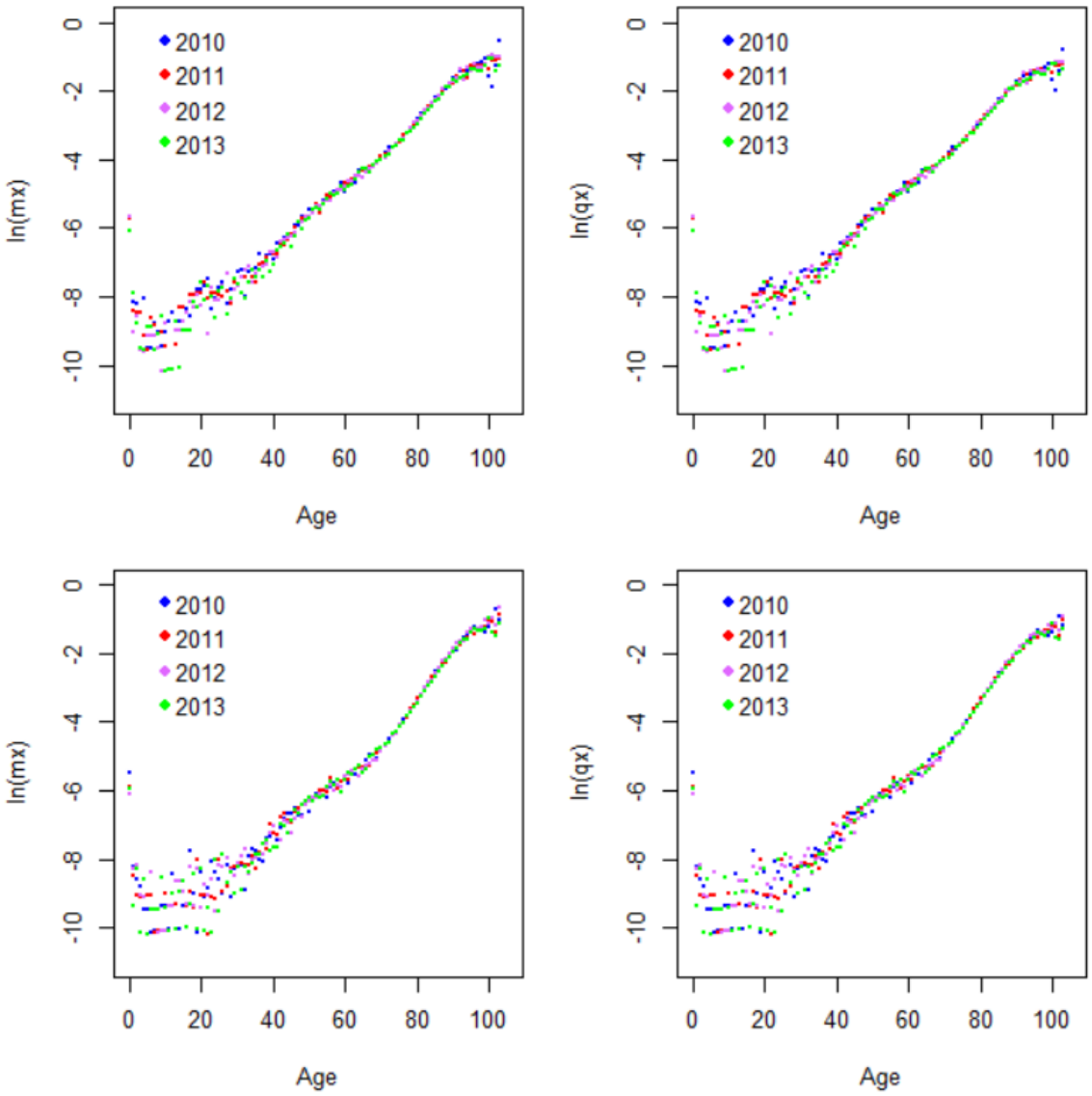

Figure 40-S. Estimated rates of death, m_x, and probabilities of death, q_x, of either CP_UD_UB (closed population and uniform distribution of deaths and births) or OP_UD_UM_UB (open population and uniform distribution of deaths, migrants and births) scenarios. Left upper panel: $\ln(m_x)$ for men. Right upper panel: $\ln(q_x)$ for men. Left lower panel: $\ln(m_x)$ for women. Right lower panel: $\ln(q_x)$ for women.

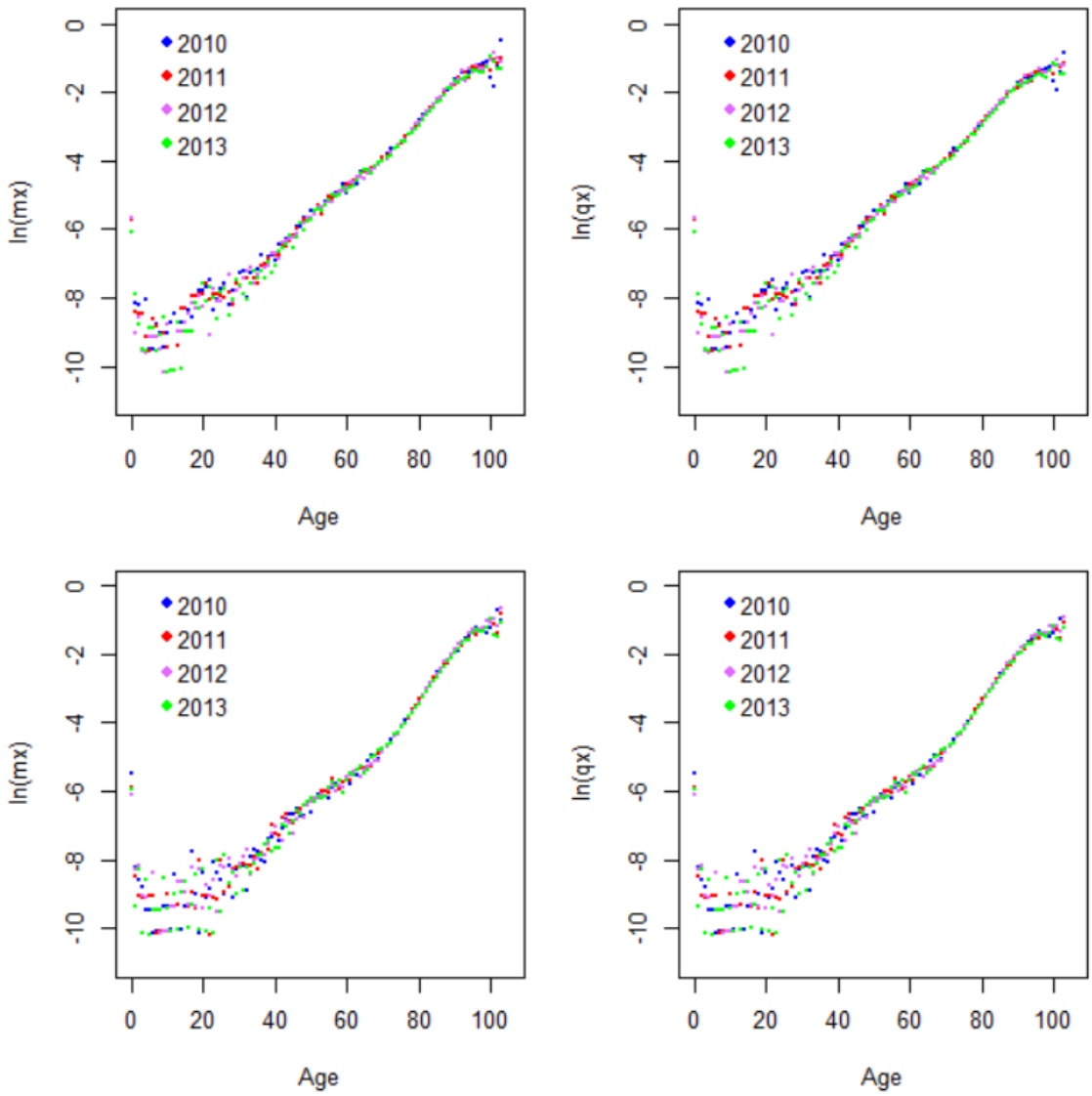

Figure 41-S. Estimated rates of death, mx, and probabilities of death, qx, of closed population with no hypothesis about distribution of deaths and uniform distribution of births (CP_NUD_UB) scenario. Left upper panel: $\ln(m_x)$ for men. Right upper panel: $\ln(q_x)$ for men. Left lower panel: $\ln(m_x)$ for women. Right lower panel: $\ln(q_x)$ for women.

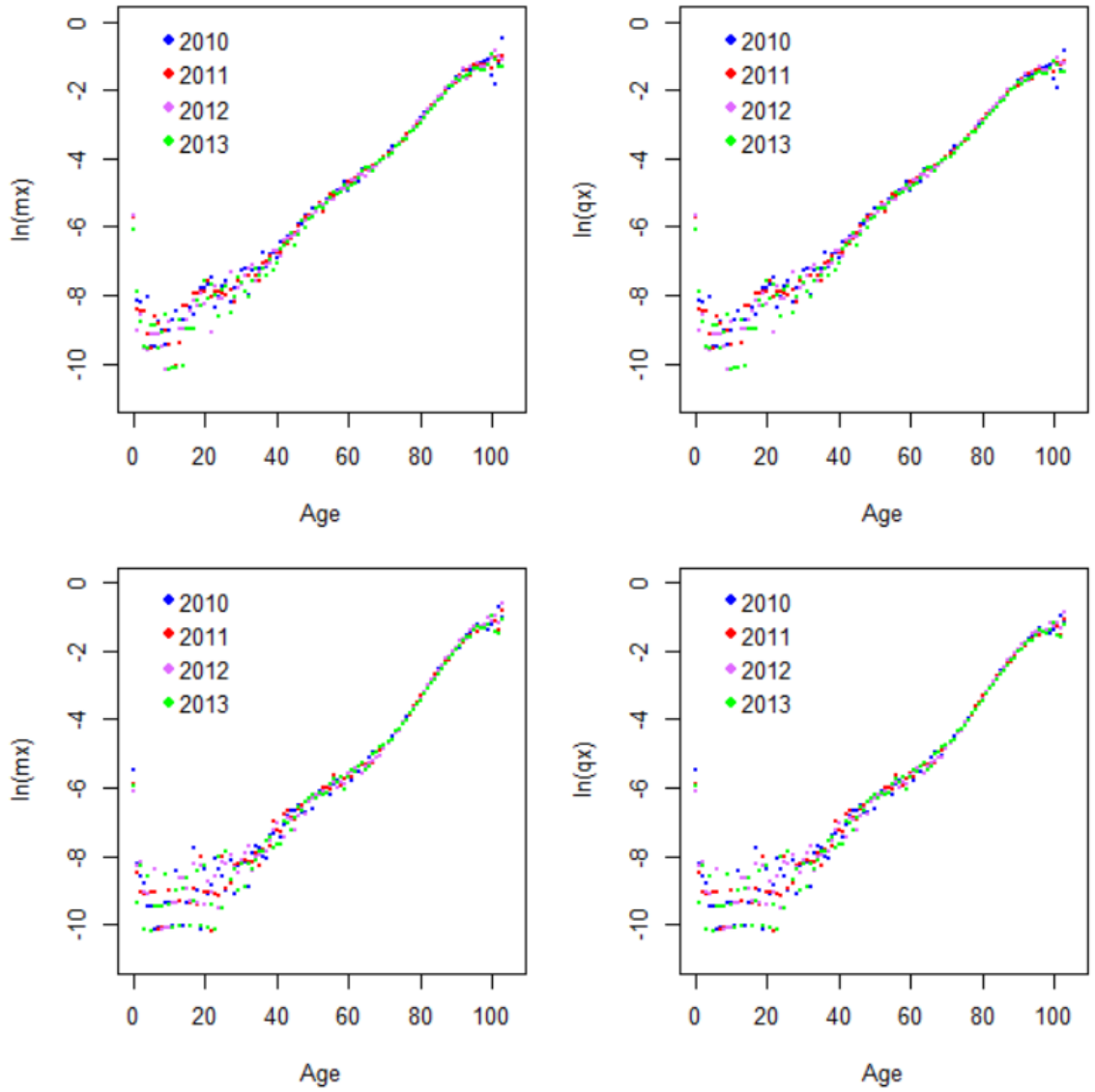

Figure 42-S. Estimated rates of death, mx, and probabilities of death, qx, of open population with no hypotheses about distribution of deaths and migrants and uniform distribution of births (OP_NUD_NUM_UB) scenario. Left upper panel: $\ln(m_x)$ for men. Right upper panel: $\ln(q_x)$ for men. Left lower panel: $\ln(m_x)$ for women. Right lower panel: $\ln(q_x)$ for women.

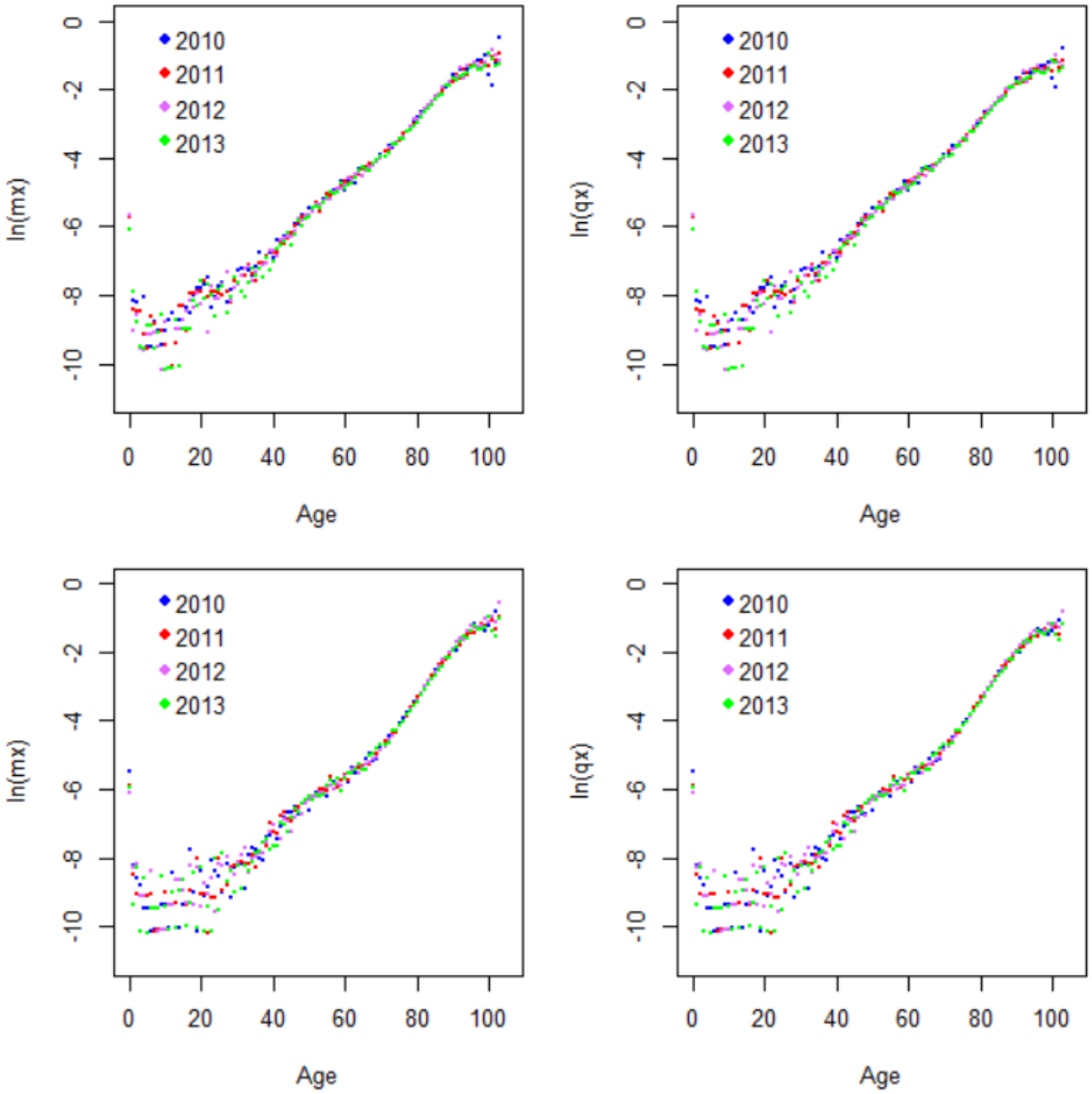

Figure 43-S. Estimated rates of death, mx, and probabilities of death, qx, of open population with no hypotheses about distribution of deaths, migrants and births (OP_NUD_NUM_NUB) scenario. Left upper panel: $\ln(m_x)$ for men. Right upper panel: $\ln(q_x)$ for men. Left lower panel: $\ln(m_x)$ for women. Right lower panel: $\ln(q_x)$ for women.

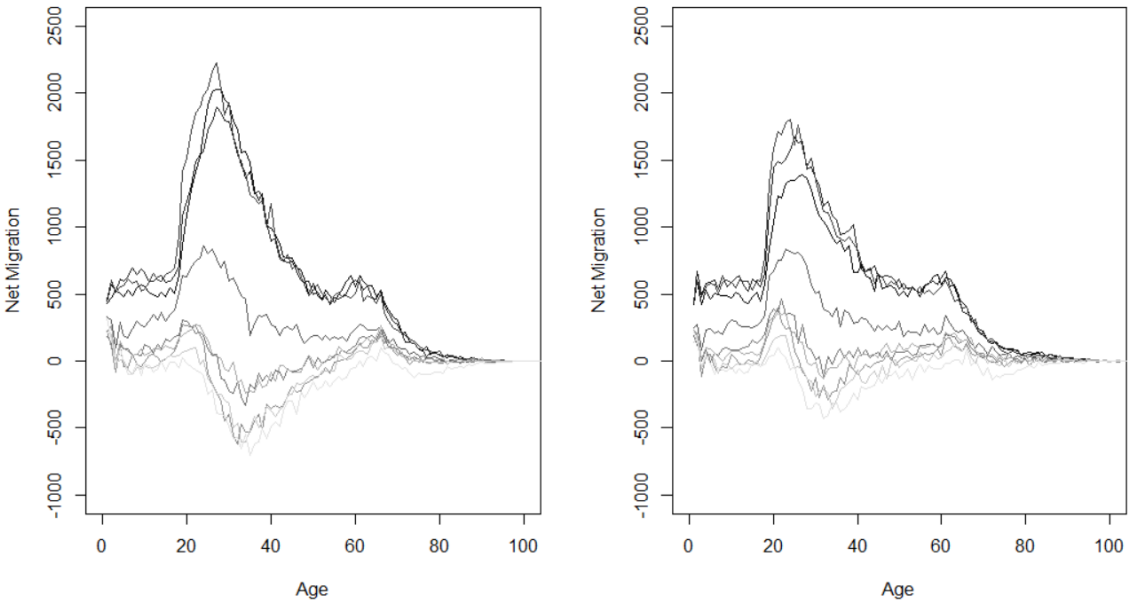

Figure 44-S. Net migration (immigrant minus emigrant), 2005-2013 years. 2005 is represented with the strongest color while 2013 is represented with the weakest color. Left panel for men and right panel for women.

Transformations in Weekly Birth Distribution. A Temporal Analysis 1940-2010 (A4)

Josep Lledó, Universitat de Valencia

Jose M. Pavía, Universitat de Valencia

Francisco G. Morillas, Universitat de Valencia

ABSTRACT

The seasonality of births is an issue that has received much attention in the literature. Empirical evidence has revealed that annual birth distributions have evolved from environmentally-regulated fertility patterns to models that are dominated by socio-cultural factors. This study expands upon this literature by examining the weekly distribution of births and its evolution over the past decades. Using the micro-data of the municipal register files of the Valencian Community (n=3,674,110), we empirically reveal that the current organization of work schedules in the healthcare sector impacts the weekly distribution of births, resulting in the hegemonic position of the medical profession. This has been confirmed by the change in the distribution of births, by days of the week, since the mid-20th century until the present date.

Keywords: Comunitat Valenciana, Births, Healthcare Organization, Healthcare Policy

Acknowledgements

The authors wish thank three anonymous reviewers for their valuable comments and suggestions, the *Instituto Valenciano de Estadística* (Valencia Institute of Statistics) and especially, Francisco Fabuel, for providing with the microdata that was used in this study and Professor Carles Simó for all of his constructive suggestions. The authors appreciate all of the support that they have received for the projects CSO2013-43054-R and MTM2016-74921-P financed by the Ministry of Economy and Competitiveness.

Introduction

The 20th century has been characterized as a period of profound social and economic changes in European countries. The emergence of social welfare systems, the massive incorporation of females in the labor market and the progressive reduction, homogenization and regulation of the workday are just some of its more noteworthy milestones (Juárez, 1993; Muñoz de Bustillo, 2003; Poal, 1993; Mósesdóttir *et al.,* 2006; Olmos and Silva, 2011). Along with a global reduction in the workday, we have also seen a concentration of work (and leisure) hours over the same times of the day, the same days of the week and the same weeks of the year (Prieto *et al.,* 2008).

Progressively, today's society has begun to take notice of the importance of time, given that this is a non-exchangeable and nonrecoverable element. Currently, the problem exists not so much in the quantity of time worked but in the organization of the same, and in its suitability to social and family needs. This harmonization and optimization of time impacts the reproductive project, especially the planning of maternity. So, the study of the seasonality of births has been an issue receiving much attention in the field of sociology. Knowing how births are distributed across the year and how their monthly distribution has changed across time has become a subject of great sociological interest given that it serves as a reflection of the changes taking place in society. In the current discussion, changes in seasonality are related with demographic changes associated with a major decrease in fertility, with parents controlling the number of children that they are going to have and when they will have them. Fertility has become a voluntary act, related to the theory of *rational choice* (Elster, 1986).

In other words, seasonality has resulted from the process in which, given access to the effective control of fertility, this may be conceptualized as a result of the parent's decision regarding: how many children to have, what time of year to have them and even, how to temporarily space out the births (Cordero, 2009). Empirical evidence reveals the evolution from natural fertility patterns, in which environmental factors play a major role in the regulation of childbearing, towards models that are defined by socio-cultural factors.

Historically, environmental conditions at the time of conception have been a major factor correlating with the seasonality of births (Fuster, 1989; Rusell *et al.,* 1993). For example, in northern Europe, an increased number of births tended to take place over the springtime (conceived in the summer), with fewer births being recorded in the fall (winter conceptions). In the south of the United States, exactly the opposite has been observed, with more births taking place in the summer-fall and the springtime registering the lowest number of births (Lam and Miron, 1994). Over recent years, a change has taken place in the factors relating to seasonality of births, evolving from a model that was regulated by the environment during the 1940-1960 period to one that is based on socio-cultural conditioners (Quesada, 2006). Specifically, mothers who are between the ages of 25 and 34, married and have a higher education level, present greater seasonality in the births of their second or third children as compared to the birth of their first and fourth child as compared to mothers under the age of 19 or over the age of 35, those who are unmarried and have low education levels (Bobak and Gjonja, 2001).

Despite this, limited empirical evidence exists regarding the weekly distribution of births and the changes taking place over recent decades. This work attempts to delve further into the seasonality of the fertility phenomenon and is novel, given that, within the tradition of studies on seasonality, it examines the distribution of births taking place between the days of the week. The most relevant issue is whether or not a weekly distribution pattern in births exists that differs between the younger and older generations, and, if it does in fact exist, if it is caused by socio-cultural factors.

Our hypothesis is that the time of birth has gone from being a purely biological act to one that is institutionalized, governed mainly by physicians, who use their hegemonic position to adjust the process according to their individual and group schedules. Based on the study of another aspect of the seasonality of births, which has yet to receive attention, the results of this study may place certain limitations on that which has been concluded until now regarding this literature and may provide a sort of break with the prevailing paradigm as they reduce the individual decision making capacity of the parents in the planning of the births.

The remainder of this article has been structured as follows: Section 2 discusses the relationship between hospital organization and the evolution of the weekly distribution of births. Section 3 describes the data and the methodological aspects. Section 4 presents the main results. Section 5 discusses and assesses the conclusions.

Births and Healthcare Organization

The socio-economic development occurring in Spain has been particularly visible in distinct areas of the social welfare system. In the field of healthcare —considered to be one of the six basic social needs (Miguel, 1996)—, and in the use resulting from the same, it has become necessary to more effectively manage available resources. The increase of the active population in the secondary and tertiary sectors, to the detriment of the primary sector, along with migrations from the country to the city and the improvement of the hygienic-healthcare and communications infrastructures (Alemany, 2014), have made possible that rising population percentages have increased their proximity to the healthcare centers and the demand of the same (Robles *et al.*, 1996). As of the 1970s, at-home labor with a midwife has (almost entirely) disappeared, progressively moving into the hospitals where medicine emerges as the main player in the exercising of its "absolute control" (Montes, 2007). So, currently, when referring to the social representations and practices of birth, we should do so based on the expert opinions proposed by the medical system.

The set of changes taking place in Spanish society have led to some major cultural transformations that have determined the appearance of new social values and behaviors, such as the cases of physicians who only attend planned labors and cesareans or even, the cases of women who, once pregnant, only wish to have a cesarean birth in order to avoid painful labor contractions (Montes, 2007).

In the Spanish healthcare system, the number of cesareans has been progressively on the rise. For example, from a cesarean rate of 22.45% in 2001, it increased to 25.20% in 2005, meaning a 12.3% increase in the proportion of cesareans in proportion to all births (Ministry of Health and Social Policy, 2009). The management of healthcare resources, the inherent risks of pregnancy (Ronda *et al.,* 2009), the fatigue of the mother, and especially, the adaptation of the labor environment have all favored, along with other social aspects (Moroto *et al.,* 2004), a more active control by healthcare

professionals at the time of birth. The precise time of birth not only depends on the random biological process but also on the programming of the same, even its induction, which is becoming more and more frequent. Hospital and interventionist labor make up an authentic cultural pattern of care, in which humanization has been diluted in the search for the complete annulment of risk (Hernández and Echevarría, 2015), and, we add, in the search for adaptation to the organization of our society's work schedule, ever more focused on the Monday to Friday timetable. In fact, the hypothesis of this study is that this responds ever more to the internal organizational of hospital centers.

A priori, it may be anticipated that if births take place without planning, the proportions of their occurrence over each of the seven days of the week should follow a discrete uniform distribution probability, with the observed discrepancies being solely attributable to the occurrence of a random phenomenon. So, when the distribution is uniform, the anticipated proportion of births for each day of the week is 14.28%.

The mass function of a discrete uniform distribution may be graphically visualized as a horizontal line. For example, Figure 1 shows the theoretical distribution (uniform) together with the distribution of births observed in the Valencia Community for the 1960-1969 decade (a period characterized by some very high birth rates and not affected by the social-economic and hospital organization changes which are the area of discussion of this study).

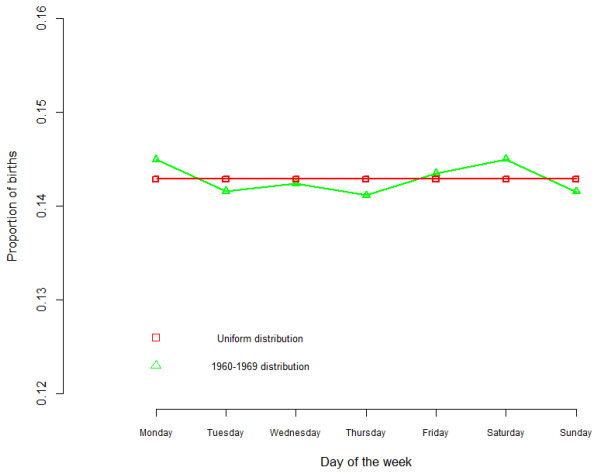

Figure 1. Distribution of births by day of the week

Data and Methodology

In order to test the hypothesis of uniform weekly distribution of births it is necessary to have adequate information regarding the dates when these births took place. In Spain, this information is included in the birth certificates. Over recent years, however, various studies have placed doubt on the reliability and completeness of the historical birth records, both in our country (Río *et al.*, 2010; Juárez *et al.*, 2012) as well as abroad (Northam and Knapp, 2016), concluding that the number of imprecisions increases as we go back in time.

With the goal of using appropriate and reliable information, this study presents a methodological innovation in which the selection and treatment of information is respected such that the sample used meets the desired characteristics both in terms of representativeness and quality. Specifically, the information used in this study includes the dates of births of the Spanish population born over the past 70 years and those who remain alive in 2010, contained in the Municipal Registers of Inhabitants of the Valencia Community from the year 2010 (n=3,674,110).

The selection of the Valencia Community comes in response to issues of opportunity. There are no reasons to believe that the changes taking place in the Valencia society in its demographic, social and healthcare dynamics differ significantly from those experienced throughout the rest of Spain. Furthermore, from an economic, demographic, social and cultural point of view, over recent decades, the Valencia Community has tended to be situated in the average of Spain. So, in our opinion, the results obtained may be applicable to Spain in general.

As for the quality of the data used, although currently the birth dates for the registration of Spaniards in the Municipal Register of Inhabitants comes from the birth certificates, this has not always been the case. Until 1996, the data appearing in the census was collected from self-reporting. The first time that a Spaniard was registered in the census, he/she introduced all of his/her information in a form, from which this data was transferred to the census file. So, even though the dates of birth appearing in the census for elderly individuals may be subject to various types of errors (including errors of memory, incorrect transfer of the exact date of birth from parents to children

or operational faults), the potential deviations that may contain the statistic shall be random and therefore, shall not systematically affect any specific day of the week

On the other hand, from a more technical point of view, the information used may be considered to be a random sample of the total of births taking place in the Community of Valencia every week during the 1940-2010 period. It is reasonable to assume, and there is no reason to suggest otherwise, that there is no relationship between the deaths of those individuals born as of 1940 and who are no longer living in 2010 and the day of the week on which they were born. The available population may be considered to be a random sample (at least in terms of the day of the week of the birth) of the register of births for each year considered.

In order to assess the hypothesis of the uniformity of births, two statistical analyses have been conducted. On the one hand, chisquare tests of goodness of fit have been conducted (DeGroot, 2003) by decade in order to test the uniformity hypothesis. In the tests, the null hypothesis checks whether or not *the probability distribution of the proportion of births by days of the week follows a uniform probability distribution;* as compared to the alternative hypothesis that *the distribution is not uniform*. On the other hand, the evolution of the daily proportion of births has been analyzed (with respect to the annual total) over the last seventy years and the data has been treated using time series analysis techniques (Uriel and Peiró, 2000). The proportion of births on each day has been calculated in relation to the total of the year in order to study the evolution of the daily proportion of births (see Figure 2 left). If we assume that there are no seasonal or calendar effects (such as weekly cycle), the distribution of the evolution of the daily number of births shall be uniform, with the anticipated birth proportions being 1/365 and 1/366 (horizontal line Figure 2-left) for, respectively, non-leap years and leap years. The ordered sequence of proportions ($p_{i\tau}$) of daily births, which we call X_t, is analyzed as a time series.

$$p_{i\tau} = \frac{n_i^\tau}{N^\tau} = \frac{Total\ of\ births\ on\ day\ i\ of\ year\ \tau}{Total\ of\ births\ in\ year\ \tau}, for \begin{Bmatrix} i = 1, \dots 365\ (366\ leap\ years) \\ \tau = 1940, \dots 2010 \end{Bmatrix}$$

Initially, when there is a single variable, it is not possible to calculate the correlation measures. However, when there is an ordering structure in the data, a new variable may be considered with the period values coming from the current X_{t-1}. So, the

autocorrelation of order one is defined as the correlation between the variables X_t and X_{t-1}. In general, the autocorrelation of order k measures the correlations between the variables X_t and X_{t-k}. In order to eliminate the correlation between two variables, the effect of the third variables obtains the partial auto-correlation. The function of the partial auto-correlation of pit is offered in Figure 2-right.

The graphic representation of the partial auto-correlation function (pacp) allows for the observation of the intensity of the relationship between certain values and their delayed homologues 1, 2, 3… time units. A larger value of the pacp indicates a higher correlation intensity. Figure 2 right clearly shows how the values separated by 7 time units (and its multiples) have a greater statistical correlation. This indicates that every day of the week has values similar to the homologue of the rest of the weeks. It is interesting to observe how the delays of order 6 and 8 (surrounding the order 7 delays) are also significant, probably due to possible deviations of the dominant pattern resulting from the annual holiday calendar.

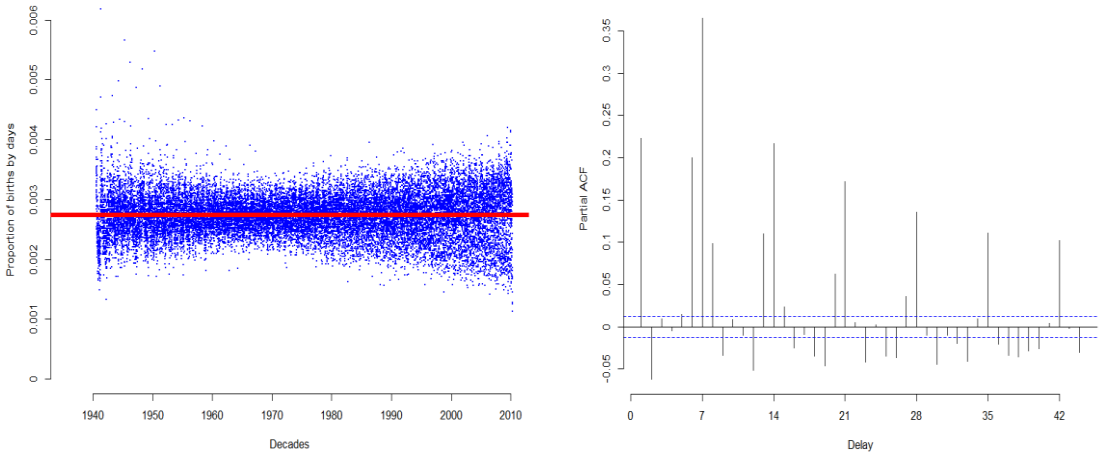

Figure 2. Distribution of births in the Valencia community between 1940 and 2010. Left panel: time series of the proportions of daily relative births, 1940-2010. Right panel: function of partial auto-correlation of the time series

Results

This section presents the results of implementing the methods described in the previous section. The results of the hypothesis tests confirm that statistically, a gradual change has taken place in the weekly distribution of births over recent decades (see Table I).

Table I. p-values of the chi-squared contrast

	1940-1949	1950-1959	1960-1969	1970-1979	1980-1989	1990-1999	2000-2009
p-values	0.07553	0.04548	0.06622	$3.341*10^{-5}$	$9.711*10^{-9}$	$4.969*10^{-14}$	0.000

Goodness of fit tests indicate that for the decades 1940-1949, 1950-1959 and 1960-1969, the null hypothesis of uniform distribution of births is not rejected, for a significance level of 1%. On the other hand, it is seen that for each of the following decades, the null hypotheses are rejected, verifying an increase in the probabilities of rejection over time: the p-values are decreasing. These results are in line with that presented in Figure 3, which shows the estimate of the weekly birth distributions for each of the seven decades.

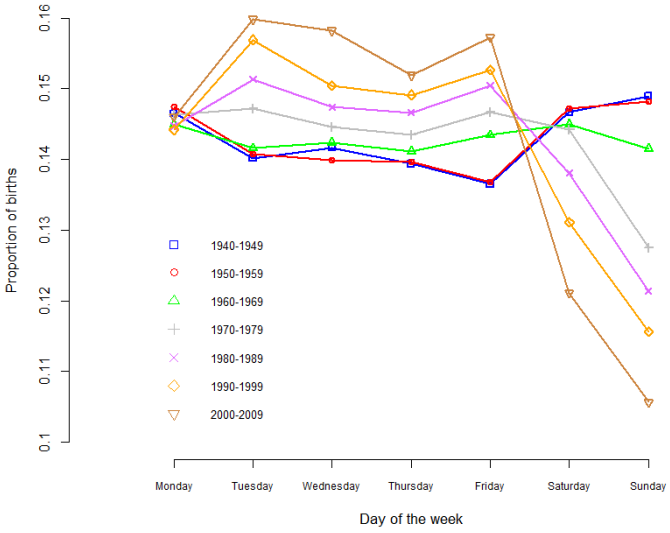

Figure 3. Distribution of births by day of the week over the last seven decades

As seen in Figure 3, an ever-increasing rise has taken place in births over the central days of the typical work week. The days having the highest number of births are Wednesdays and especially, Tuesdays; which represented 14.71% of all births of the week in the 1970-1979 period as opposed to 16.98% in the 2000-2009 period. Since the 1970s, a major decrease has also been observed in the births taking place on Sundays (14.15% in the 1960-1969 period as compared to 10.56% in the 2000-2010 period), which over the past two decades, has also extended, with increasing intensity, to Saturdays.

To complete the study, a behavior analysis was conducted for the time series of births. The order 7 delays and, to a lesser extent, those of order 6 and 8 (Figure 2-right), indicate a strong relationship between the proportions of births every 7 days. This

reinforces the idea of the existence of a clear weekly cycle. On the other hand, the order 2 delays reveal a negative dependence on the proportion of births every 2 days, which is consistent with assuming that those births taking place in advance (or delayed) do not take place (or take place) over the subsequent days. Furthermore, it has been proven that the programming of the day of the week of the births has been accentuated over recent decades. The partial auto-correlation of the order 7 delay is more pronounced when considering only the values corresponding to the last three decades of the time series and even more so for the last decade.

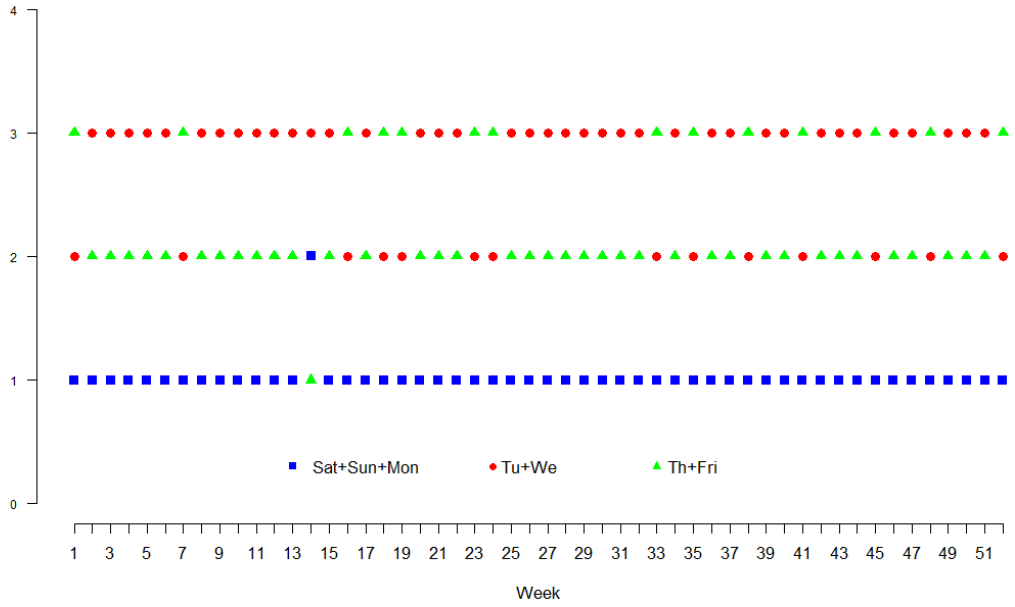

Figure 4. Ordering of the normalized proportion of births by day of week. Year 2007

To delve further into the causes of the results shown in Figure 3 and in order to study the effect of the yearly holidays on the weekly cycle, Figure 4 has been created, in which it is observed, in an ordered manner and grouped by blocks of days, the days having (relatively) more births for each week of the year 2007[39]. In Figure 4, we can observe that in the majority of the weeks of the year, the central days (Tuesday and Wednesday) have a greater proportion of births. Only on specific dates (Christmas, Easter Week and summer holidays) do we find that the overall pattern of a greater number of births on Tuesdays and Wednesdays does not take place, with a greater effect existing for the specific days of the week on which the holiday falls. We observe that the

[39] The selection of the year 2007 has been made because in this year, Christmas day falls on a Tuesday and because it corresponds to one of the last years analyzed therefore the effects of healthcare organization should be more intense.

births on the Thursday and Friday of Easter Week move to last place during this holiday week (week 14).

To summarize, the statistical analyses carried out in this study reinforce the initial hypothesis that both the mentioned sociocultural and socio-labor factors have modified the weekly distribution of births so as to create a weekly cycle structure that is more adapted to the typical work week.

Conclusions

The relevance of this work lies in the progressive emphasis placed on the *time* variable in the social sciences scenario. For the first time in the literature, it has been clearly shown how the way the Spanish society organizes its times of work and leisure, especially in the healthcare sphere, is significantly modifying the weekly distribution of births in the country. It has been revealed that the weekly distribution of births has evolved from a *coherent* uniformity in the year 1970 and previous ones, towards a weekly cycle structure, more adapted to the typical work week, in which aspects of social organization come into play.

The hypothesis of (discrete) uniform distribution for weekly births has been tested in this study by using the available data in an imaginative manner. Given the difficulty of obtaining detailed and reliable birth statistics, especially for the mid-20th century, figures were used from the population registers of 2010 for the Valencia community (which only considers individuals who are living in this year).

The results of this work reveal major differences in the proportion of births over each of the days of the week. In the mid-20th century, the proportion of births over each of the days was similar to a uniform probability distribution. However, due to social changes taking place over the final decades and which favored an improved management and technical competency by the healthcare services (Moroto *et al.,* 2004), it has evolved (Figure 3) towards a situation in which there are ever more programmed births during the central days of the week as compared to the weekends.

One of the strengths of the study is that after treating the proportion of births as a time series, the partial auto-correlation function shows a clear dependence on data separated by multiples of 7. The strong autocorrelation every 7 days suggests that the

results that were observed, regarding a concentration of births over the central days of the week, is becoming structural, not corresponding to a spurious phenomenon.

Furthermore, each proportion of births by day of the week that were distinct from 14.28% (uniform distribution) is artificially caused by human intervention. During the first decade of the 21st century, a reduction of approximately 25% was observed in the number of births that randomly corresponded to Sundays. On the other hand, the percentage of births occurring during the central days of the work week (Tuesdays and Wednesdays) are higher than expected for the case of a uniform distribution. The number of births during these days is approximately 10% greater than what would be expected to randomly occur.

In addition to determining how social organizations (as well as the power relationship between physician-patient) may have serious impacts on the biological variables, the results of this study also have implications on the actuarial and demographic fields. The analysis of the time distribution of birth for a population is relevant for the creation of mortality or life tables (Pavía *et al.,* 2012; Lledó *et al.,* 2016). Life tables measure the incidence of mortality for the population residing in a country for one year for each of the ages. One of the hypotheses used for its creation is to assume a uniform distribution of the days of birth for all of the individuals in the population who did not pass away during said year (INE, 2015). When this hypothesis is not fulfilled, as tested in this study, the estimates obtained are inefficient, perhaps having an impact that should be examined, on all of those areas where the mortality table is used: the calculation of life expectancy, the determination of insurance premiums or the estimation of future pensions.

Likewise, the increase in programmed births may also introduce unexpected disruptions in the samples constructed for study in the social sciences and in clinical trials. Given that as an experimental control, *date of birth* is one of the most commonly used variables in the systematic assignment of individuals to groups (Idoate and Idoipe, 2002), subjects born on the days with unplanned births have a greater probability of being selected. This may have some relationship to the style of life of the parents and may surreptitiously introduce a variable of confusion (Halperin and Heath, 2012) that influences the experiment's results.

To summarize, this study contributes to the literature examining seasonality of births and its changes. On the one hand, the weekly cycle is examined, an aspect that has received little attention in the literature. On the other hand, certain limitations are found in the individual decision making capacity of the parents, upon observing that the time of birth adjusts more to the individual needs and desires of the healthcare professionals. In this line, it may be useful to study whether or not, along with the concentration of births taking place over the workdays, there has also been an accumulation of the same during specific hours of the day. The time of birth has become a socially controlled event having potential repercussions in different areas and fields of action such as demography, management of healthcare resources and even experimentation in the social sciences.

Bibliografía

ABS (2016). *Life Tables, States, Territories and Australia, 2012-2014 (Explanatory Notes)*, Australian Bureau of Statistics. Available at goo.gl/Qodth2.

Ahrems, W., and Pigeot, I. (2007). *Handbook of Epidemiology.* NewYork: Springer.

Alemany, M. J. (2014). *Matronas y cambio social en la segunda mitad del siglo XX. De mujeres y partos*. Tesis doctoral, Universitat de Valencia.

Arias, E., Rostron, B. L., and Tejada-Vera, B. (2010). United States Life Tables, 2005, *National Vital Statistics Reports*, 58, 1-132. Avalilable at goo.gl/0jSJRt.

Arias, E. (2015). United States Life Tables, 2011, *National Vital Statistics Reports*, 64, 1-64. Avalilable at goo.gl/BOkbrT.

Ayuso, M., Corrales, H., Guillen, M., Pérez-Marín, A.M., y Rojo, J.L. (2007). *Estadística Actuarial Vida*. Barcelona: Publicacions Universitat de Barcelona.

Baddeley, A., and Turner, R. (2005). An R Package for Analyzing Spatial Point Patterns, *Journal of Statistical Software,* 12, 1-42.

Baddeley, A, Diggle, P.J., Hardegen, A., Lawrence, T., Milne, R.K., and Nair, G. (2014). On tests of spatial pattern based on simulation envelopes, *Ecological Monographs*, 83, 447-489.

Barrieu, P., Bensusan, H., El Korouis, N., Hillairet, K., Loisel, S., Ravanelli, C., and Salhi, Y. (2012). Understanding, Modelling and Managing Longevity Risk: Key Issues and Main Challenges, *Scandinavian Actuarial Journal*, 3, 203-231, DOI: 10.1080/03461238.2010.511034.

Basulto, J., y García, J. (2009). *Historia de la Probabilidad y la Estadística (IV)*, Huelva-Sevilla: Universidad de Huelva.

Benjamin, B., and Pollard, J. (1986). *The Analisys of Mortality and Other Actuarial Statistics*, London: Heinemann.

Benjamin, B., and Pollard, J. (1992). *The Analysis of Mortality and Other Actuarial Statistics*. London: Butterworth-Heinemann.

Berko, J., Ingram, D. D., Saha, S., and Parker, J. D. (2014). Deaths Attributed to Heat, Cold, and Other Weather Events in the United States, 2006-2010, *National Health Statistics Report*, 76, 1-5.

Biffis, E., and Blake, D. (2009). Mortality-linked Securities and Derivatives. Discussion Paper PI-0901, Pensions Institute, London, UK, DOI: 10.2139/ssrn.1340409.

Blake, D., Cairns, A., and Dowd, K. (2006). Living with Mortality: Longevity Bonds and Other Mortality-linked Securities, *British Actuarial Journal*, 12, 153-228, DOI: 10.1017/s1357321700004736.

Bobak, M., and Gjonca, A. (2001). The Seasonality of Live Birth is Strongly Influenced by Socio-demographic Factors, *Human Reproduction*, 16, 1512-1517, DOI: 10.1017/S0021932000018204.

Booth, H., and Tickle, L. (2008). Mortality Modelling and Forecasting: A Review of Methods, *Annals of Actuarial Science*, 3, 3-43.

Börger, M., Fleischer, D., and Kuksin, N. (2014). Modeling the Mortality Trend under Modern Solvency Regimes, *ASTIN Bulletin*, 44, 1-38. DOI: 10.1017/asb.2013.24.

Brasche, O. (1870). *Beitrag zur Methode der Sterblichkeitsberechnung und zur Mortalitätsstatistik Russland's*, Würzburg: A. Struber's Buchhandlung.

Cabrer, B., y Pavía, J.M. (2003). Flujos Demográficos Regionales. Un Análisis Input-Output, *Revista Estadística Española*, 45, 407-429.

Cairns, A., Blake, D., and Dowd, K. (2008). Modelling and management of mortality risk: A Review, *Scandinavian Actuarial Journal*, 2, 79-113.

Cairns, A., Blake, D., Dowd, K., Coughlan, G. D., and Khalaf-Allah, M. (2011). Bayesian Stochastic Mortality Modelling for Two Populations, *ASTIN Bulletin*, 41, 29-59, DOI: 10.2143/AST.41.1.2084385.

Cairns, A. (2013). Robust hedging of longevity risk, *The Journal of Risk and Insurance*, 80, 621-648.

Cairns, A., Blake, D., Dowd, K., and Kessler, A. R. (2016). Phantoms Never Die: Living with Unreliable Population Data, *Journal of Royal Statistical Society, Series A*, 179, 1-31, DOI: 10.2139/ssrn.2676648.

Carstensen, B. (2007). Age-period-cohort Models for the Lexis Diagram. *Statistics in Medicine*, 26, 3018-3045, DOI: 10.1002/sim.2764.

Caselli, G., Vallin, J., and Wunsch, G. (2006). *Demography: Analysis and Synthesis. A Treatise in Population Studies*, Burlington, MA: Elseiver.

Coale, A.J., and Demeny, P. (1966). *Regional Model Life Tables and Stable Populations*, Princeton: Princeton University Press.

Conover, W. J. (1971). *Practical Nonparametric Statistics*, New York: John Wiley and Sons.

Copas, J. B., and Haberman, S. (1983). Non parametric graduation using kernel methods. *Journal of the Institute of Actuaries*, 110, 135-156.

Cordero, J. (2009). El espaciamiento de los nacimientos: una estrategia para conciliar trabajo y familia en España, *Revista Española de Investigaciones Sociológicas*, 128, 11-33.

Cressie, N. (1991). *Statistics for Spatial Data*. New York: John Wiley and Sons.

Cummins, J. D., and Santomero, A. M. (2005). *Changes in the Life Insurance Industry: Efficiency, Technology, and Risk Management*. Innovations in Financial Markets and Institutions.

Currie, I.D. (2016). On fitting generalized linear and non-linear models of mortality, *Scandinavian Actuarial Journal*, 356-383.

Danesi, I.L., Haberman, S., and Millossovich, P. (2015). Forecasting mortality in subpopulations using Lee-Carter type models: a comparison, *Insurance: Mathematics and Economics*, 62, 151–161.

DeGroot, M. H. (2003). *Probabilidad y Estadística*. Addison-Wesley Iberoamérica.

Deschenes, O., and Moretti, E. (2007). Extreme Weather Events, Mortality and Migration, Working Paper No. 13227, National Bureau of Economic Research, DOI: 10.3386/w13227.

Ediev, D., Coleman, D., and Scherbov, S. (2014). New Measures of Population Reproduction for an Era of High Migration, *Population, Space and Place*, 20, 622-645.

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, New York: Chapman and Hall.

Elster, J. (1986). *Rational Choice*. Oxford: Basil Blackwell.

Enchev, V., Klinow, T., and Cairns, A. (2016). Multi-population mortality models: fitting, forecasting and comparisons, *Scandinavian Actuarial Journal* [online], DOI: 10.1080/03461238.2015.1133450.

Esipova, N., Ray, J., and Publieses, A. (2001). Gallup World Poll: The Many Faces of Global Migration, *IOM Migration Researh Series*, 43, 5-71.

Eurostat (2010). Work Session on Demographic Projections, Lisbon 28-30 April 2010, Eurostat Methodologies and Working Papers, Luxembourg: Publications Office of the European Union.

Fernández, J. J., and Gregorio, M. M. (2015). Seasonal Mortality for Fractional Ages in Short term Life Insurance, *Scandinavian Actuarial Journal*, 2015, 266-277, DOI: 10.1080/03461238.2013.819028.

Forfar, D. O., MacCutcheon, J. J., and Wilkie A. D. (1988). On graduation by mathematical formula, *Journal of the Institute of Actuaries*, 115, 1-149.

Forsyth, C.H. (1914). American Life Tables, *Quarterly Publications of the American Statistical Association*, 14, 228-235, DOI: 10.2307/2964953.

Fuster, V. (1989). Seasonality of births and family characteristics in a Spanish population, *Journal of Biosocial Science*, 21, 465-474.

Gavrilov, L. A., and Gavrilova, N. S. (2011). Mortality Measurement at Advanced Ages. A Study of the Social Security Administration Death Master File, *North American Actuarial Journal*, 15, 432-447, DOI: 10.1080/10920277.2011.10597629.

Goerlich, F.J. (2008). Las Tablas de Mortalidad del Instituto Nacional de Estadística: 1900-1901 a 2004-2005. Recopilación Crítica, *Estadística Española*, 50, 523-569.

Gompertz, B. (1825). On the Nature of the Function of the Law of Human Mortality on a New Mode of Determining the Value of Life Contingencies, *Transactions of the Royal Society,* 115, 513-585, DOI: 10.1098/rstl.1825.0026.

Grabarnik, P., and Chiu, S. (2002). Goodness-of-fit test for complete spatial randomness against mixtures of regular and clustered spatial point processes, *Biometrika*, 89, 411-421. doi: 10.1093/biomet/89.2.411

Graunt, J. (1662). *Natural and Political Observations Made upon the Bills of Mortality*, London: Roycroft.

Halperin, S., and Heath, O. (2012). *Researching Politics: Methods and Practical Skills*. Oxford: Oxford University Press.

Hernández, J. M., y Echevarría, P. (2015). El nacimiento hospitalario e intervencionista: un rito de paso hacia la maternidad. *Revista de Antropología Iberoamericana*, 10, 401-426

Hinde, A. (1998). *Demographic Methods*, London: Arnold.

IABE (2015). *BIG DATA: An Actuarial Perspective,* Brussels: Institute of Actuaries in Belgium. Available at goo.gl/fAOm7x.

Idoate, A., y Idoipe, Á. (2002). Investigación y ensayos clínicos. En: J. Bonal J (ed.), *Farmacia Hospitalaria*. Madrid: Fundación Española de Farmacia Hospitalaria.

INE (2007). *Metodología Empleada en el Cálculo de las Tablas de Mortalidad de la Población de España 1992-2005*, Madrid: Instituto Nacional de Estadística. Disponible: goo.gl/6Yc0Hn.

INE (2009). *Tablas de Mortalidad. Metodología*, Madrid: Instituto Nacional de Estadística. Disponible: goo.gl/jLS2op.

INE (2010a). *Movimiento Natural de Población. Defunciones (cifras anuales).* http://www.ine.es/jaxiBD/tabla.do?per=12&type=db&divi=MNP&idtab=49&L=0

INE (2010b). *Estadística de Variaciones Residenciales (01/02/2011).* http://www.ine.es/prodyser/micro varires.htm

INE (2012). *Estimaciones de la Población Actual de España*. http://www.ine.es/jaxiBD/menu.do?L=1&divi=EPOB&his=0&type=db

INE (2015). *Tablas de mortalidad. Metodología*. Madrid: Instituto Nacional de Estadística. http://www.ine.es/metodologia/t20/t2020319a.pdf.

INE (2016). *Tablas de Mortalidad*, Madrid: Instituto Nacional de Estadística. Disponible: goo.gl/8Ywdc9.

Juárez, M. (1993). La cultura del ocio y su función de cambio social hacia el final del siglo XX, *Revista Complutense de Educación*, 4, 29-25.

Juárez, S., Alonso, T., Ramiro-Fariñas, D., and Bolúmar, F. (2012). The Quality of Vital Statistics for Studying Perinatal Health: the Spanish Case, *Pediatric and Perinatal Epidemiology*, 26, 310-5.

Kelly, J.J. (1987). Improving the comparability of international migration statistics: contributions by the Conference of European Statisticians from 1971 to date, *The International Migration Review*, 21, 1017-1037.

Khoo, S.E., and McDonald, P. (2011). The Demographic Dynamics of Migration Processes, Working Paper for the Department of Immigration and Citizenship, Australian Demographic and Social Researh Institute. Available at goo.gl/ChmzOG.

Lam, D. A., and Miron, J. A. (1994) .Global Patterns of Seasonal Variation in Reproductive Outcomes, *Annals of the New York Academy of Sciences*, 709, 9-28. DOI: 10.1111/j.1749-6632.1994.tb30385.x.

Lazar, D., and Denuit, M.M. (2009). A multivariate time series approach to projected life tables, *Applied Stochastic Models in Business and Industry*, 25, 806-823.

Lee, R. D., and Carter, L. R. (1992). Modelling and Forecasting U.S. Mortality, *Journal of the American Statistical Association*, 87, 659-671. DOI: 10.1080/01621459.1992.10475265.

Lee, R., and Miller, T. (2001). Evaluating the Performance of the Lee-Carter Method for Forecasting Mortality, *Demography*, 38, 537-549, DOI: 10.2307/3088317.

Lexis, W. (1880). La Représentation Graphique de la Mortalité au Moyen des Points Mortuaires. *Annales de Démographie Internationale*, 4, 297-324.

Li, J. S., and Hardy, M. R. (2011). Measuring Basis Risk in Longevity Hedges, *North American Actuarial Journal*, 15, 177-200, DOI: 10.1080/10920277.2011.10597616.

Li, N., and Lee, R. (2005). Coherent mortality forecasts for a group of populations: an extension of the Lee-Carter method, *Demography*, 42, 575-594.

Li, O., and Racine, J.F. (2007). *Nonparametric Econometrics*. New Jersey: Princeton University Press.

Livi Bacci, M. (2000). *Introducción a la Demografía*. Barcelona: Ariel.

Loosmore, N. B., and Ford, E. D. (2006). Statistical Inference using the g or K Point Pattern Spatial Statistics, *Ecology,* 87, 1925-1931. DOI: 10.1890/0012-9658(2006)87[1925:SIUTGO]2.0.CO;2.

Lledó, J., Pavía, J. M., and Morillas, F. (2016). Assessing Implicit Hypotheses in Life Table Construction, *Scandinavian Actuarial Journal*, DOI: 10.1080/03461238.2016.1177585.

Makeham, W.M. (1860). On the Law of Mortality and the Construction of Annuity Tables, *Journal of the Institute of Actuaries and Assurance Magazine*, 8, 301-310, DOI: 10.1017/s204616580000126x.

Massey, D. S., Aragon, J., Huho, G., Kouaouci, A., Pellegrino, A., and Taylor, J. E. (1993). Theories of International Migration: A Review and Appraisal, *Population and Development Review*, 19, 431-466, DOI: 10.2307/2938462.

Miguel, J. M. (1996). ¿Desarrollo o desigualdad? Análisis de una polémica sociológica de medio siglo en España, *Revista Española de Investigaciones Sociológicas*, 75, 55-108.

Ministerio de Sanidad y Política Social (2009). *Maternidad Hospitalaria. Estándares y Recomendaciones*. Madrid.

Montes, M. J. (2007). *Las culturas del nacimiento*. Tesis doctoral, Universitat Rovira i Virgili.

Moroto, G., García, M., y Mateo, I. (2004). El reto de la maternidad en España: dificultades sociales y sanitarias, *Gaceta Sanitaria*, 18, 13-23.

Mósesdóttir, L., Serrano, A., and Remery C. (2006). *Moving Europe towards the knowledge-based society and gender equality*. Bruselas: ETUI.

Muñoz de Bustillo, R. (2003). *Nuevos tiempos de actividad y empleo*. Madrid: Ministerio de Asuntos Sociales.

Muriel de la Riva, S., Cantalapiedra, M., and López, F. (2010). Towards Advanced Methods for Computing Life Tables, Working Papers 04/2010. Madrid: Instituto Nacional de Estadística.

Northam, S. and Knapp, T. (2016). The Reliability and Validity of Birth Certificates, *Journal of Obstetric, Gynecologic and Neonatal Nursing*, 1, 3-12.

Olmos, C., y Silva, R. (2011). El desarrollo del Estado de bienestar en los países capitalistas avanzados: Un enfoque socio-histórico, *Revista Sociedad y Equidad*, 1, 1-8.

ONS (2010). Mid-year Population Estimates Short Methods Guide, Hampshire: Office for National Statistics (UK). Available at goo.gl/s7oa0J.

ONS (2012). Guide to Calculating Interim Life Tables, Hampshire: Office for National Statistics (UK). Available at goo.gl/xFz7bV.

Pavía, J. M., y Escuder, R. (2003). El proceso estocástico de muerte. Diferentes estrategias para la elaboración de tablas recargadas. Análisis de sensibilidad. *Estadística Española*, 45, 243-274.

Pavía, J. M. (2011). *101 Ejercicios Resueltos de Estadística Actuarial Vida*. Madrid: Garceta.

Pavía, J. M., Morillas, F., and Lledó, J. (2012). Introducing Migratory Flows in Life Table Construction, *Statistics and Operational Research Transactions (SORT)*, 36, 103-114.

Pavía, J. M. (2015). Testing Goodness-of-fit with the Kernel Density Estimator: GoFKernel, *Journal of Statistical Software*, 66, 1-27, DOI: 10.18637/jss.v066.c01.

Pitacco, E., Denuit, M., Haberman, S., and Oliveiri, A. (2009). *Modelling Longevity Dynamics for Pensions and Annuity Business*, Oxford: University Press.

Poal, G. (1993). *Entrar, quedarse, avanzar. Aspectos psicosociales de la relación mujer-mundo laboral*. Madrid: Siglo XXI.

Posada, M., Martín, C., Ramírez, A., Villaverde, A., y Abaitua, I. (2008). Enfermedades raras. Concepto, epidemiología y situación actual en España, *Anales del Sistema Sanitario de Navarra*, 31, 9-20. DOI: 10.4321/S1137-66272008000400002

Preston, S. M., Heuveline, P., and Guillot, M. (2001). *Demography, Measuring and Modeling Population Processes*, Oxford: Blackwell Publishers.

Prieto, C., Ramos, R., y Callejo, J. (2008). Nuevos tiempos del trabajo. Entre la flexibilidad de las empresas y las relaciones de género, *Centro de Investigaciones Sociológicas*.

Quesada, A. (2006). Cambios en la estacionalidad de los nacimientos en Andalucía, España, entre 1941 y 2000. *Boletín de la Real Sociedad Española de Historia Natural*, 101, 77-85.

R Core Team (2013). R: *A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. Available at http://www.R-project.org/.

R Core Team (2016). R: A Language and Environment for Statistical Computing, Vienna: R Foundation for Statistical Computing. Available at: http://www.R-project.org/

Río, I., Castelló, A., Jané, M., Prats, R., Barona, C., Más, R., Rebagliato, M., Zurriaga, O., y Bolúmar, F. (2010). Calidad de los datos utilizados para el cálculo de indicadores de salud reproductiva y perinatal en población autóctona e inmigrante. *La Gaceta Sanitaria*, 24(2): 172–177.

Ripley, B.D. (1977). Modelling Spatial Patterns, *Journal of the Royal Statistical Society, Series B*, 39, 172-212.

Ripley, B.D. (1981). *Spatial Statistics*. Hoboken, New Jersey: John Wiley and Sons.

Robles, E., García, F., y Bernabéu, J. (1996). La transición sanitaria en España desde 1900 a 1990, *Revista Española de Salud Pública*, 70, 221-233

Rodríguez, J.M., Albarrán, I., Ariza, F., Cóbreces, V.M., and Durbán, M.L. (2015). *The Risk of Longevity and its Practical Application to Solvency II*. Madrid, Spain: MAPFRE Foundation.

Ronda, E., Hernández, A., García, A. M., y Regidor, E. (2009). Ocupación materna, duración de la gestación y bajo peso al nacimiento, *Gaceta Sanitaria*, 23, 179-185.

Ruggles, S. (2014). Big Microdata for Population Research, *Demography*, 51, 287-297, DOI: 10.1007/s13524-013-0240-2.

Rusell, D., Douglas, A., and Allan, T. (1993). Changing seasonality of birth-a posible environmental effect, *Journal of Epidemiology and Community Health,* 47, 362-367

Russolillo, M., Giordano, G., and Haberman, S. (2011). Extending the Lee-Carter model: a three-way decomposition, *Scandinavian Actuarial Journal,* 2, 96-117.

Society of Actuaries (2005). *Living to 100 and Beyond Monograph*. Orlando, Florida: e-Proceedings of the Living to 100 and Beyond Symposium.

Soneji, S., and King, G. (2012). Statistical Security for Social Security, *Demography*, 49, 1037-1060.

Tabeau, E., Van Den Berg Jeths, A., and Heathcote, C. (2001). *Forecasting Mortality in Developed Countries: Insights from a Statistical, Demographic and Epidemiological Perspective*, Kluwer Academic Publishers, Dordrecht.

Uriel, E., y Peiró, A. (2000). *Introducción al análisis de series temporales*. Valencia: Alfa Centauro.

Vandesrchirk, C. (2001). The Lexis Diagram, a Misnomer, *Demographic Research,* 4, 97-124, DOI: 10.4054/demres.2001.4.3.

Vallin, J. (1973). *La mortalité par génération en France, depuis 1899*, Institut National d´Études Démographiques, Cahier 63. Paris: Presses Universitaires de France.

Wilmoth, J. R., Andreev, K. F., Jdanov, D. A., and Glei, D. A. (2007). Methods Protocol for the Human Mortality Database. *Human Mortality Database,* University of

California Berkeley and Max Planck Institute for Demographic Research. Available at http://www.mortality.org/.

Willets, R.C. (2004). The cohort effect: insights and explanations, *British Actuarial Journal*, 10, 833-877.

Wiśniowski, A., Forster, J. J., Smith, P. W. F., Bijak, J., and Raymer, J. (2016). Integrated Modelling of Age and Sex Patterns of European Migration, *Journal of the Royal Statistical Society, Series A*, 179, 1007-1024, DOI: 10.1111/rssa.12177.

Woo, G., Martin, C.J., Hornsby, C., and Coburn, A. (2009). Prospective longevity risk analysis, *British Actuarial Journal*, 15, 235–247.